ECOGRAPHY 28: 815–829, 2005

miniREVIEW

Minireviews provides an opportunity to summarize existing knowledge of selected ecological areas, with special emphasis on current topics where rapid and significant advances are occurring. Reviews should be concise and not too wide-ranging. All key references should be cited. A summary is required.

The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance

Bruno A. Walther and Joslin L. Moore

Walther, B. A. and Moore, J. L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance – Ecography 28: 815–829.

The purpose of this review is to clarify the concepts of bias, precision and accuracy as they are commonly defined in the biostatistical literature, with our focus on the use of these concepts in quantitatively testing the performance of point estimators (specifically species richness estimators). We first describe the general concepts underlying bias, precision and accuracy, and then describe a number of commonly used unscaled and scaled performance measures of bias, precision and accuracy (e.g. mean error, variance, standard deviation, mean square error, root mean square error, mean absolute error, and all their scaled counterparts) which may be used to evaluate estimator performance. We also provide mathematical formulas and a worked example for most performance measures. Since every measure of estimator performance should be viewed as suggestive, not prescriptive, we also mention several other performance measures that have been used by biostatisticians or ecologists. We then outline several guidelines of how to test the performance of species richness estimators: the detailed description of data simulation models and resampling schemes, the use of real and simulated data sets on as many different estimators as possible, mathematical expressions for all estimators and performance measures, and the presentation of results for each scaled performance measure in numerical tables with increasing levels of sampling effort. We finish with a literature review of promising new research related to species richness estimation, and summarize the results of 14 studies that compared estimator performance, which confirm that with most data sets, non-parametric estimators (mostly the Chao and jackknife estimators) perform better than other estimators, e.g. curve models or fitting species-abundance distributions.

B. A. Walther (bawalther@zmuc.ku.dk), Zool. Museum, Univ. of Copenhagen, Universitetsparken 15, DK-2100 København Ø, Denmark (present address of B. A. W.: Centre of Excellence for Invasion Biology (CIB), Univ. of Stellenbosch, Private Bag X1, Matieland 7602, South Africa). – J. L. Moore, Conservation Biology, Zoology Dept, 15 Downing St., Cambridge CB2 3EJ, UK.

The purpose of this review is to clarify the concepts of bias, precision, and accuracy as they are commonly defined in the biostatistical literature. The statistical concepts of bias, precision and accuracy arise in situations involving measurement, sampling, and estimation. We will mention the first two situations in passing, but will mostly focus on the problem of estimation as we intend to use these definitions to evaluate the perfor-

Accepted 12 July 2005

ISSN 0906-7590

mance of statistical estimation methods (also called estimators). For example, species richness estimators try to estimate the true or total species richness from an incomplete sample of a biological community. We will use species richness estimators as an example to show how to evaluate the performance of estimators according to their bias, precision and accuracy given some sample data. Species richness estimators belong to the class of point estimators which are trying to estimate the exact value of some population parameter (e.g. species richness, or population size). Interval estimators, on the other hand, try to estimate confidence intervals for a parameter. We do not deal with interval estimators here, but mention them briefly below.

Given some sample data, a point estimator will yield some estimate of the parameter. For example, Fig. 1 shows an estimate (open circle) that misses the true value of zero by five units. With just a single estimate, it is impossible to tell what the reason for this error is. However, if further estimates (filled circles) are calculated using different sample data, we may hypothesize the main cause for the error of the first data point. Figure 1a illustrates the case when systematic error is the main cause while Figure 1b illustrates the case when random error is the main cause. In this case, we assume that both types of error are due to the estimator itself. Of course, systematic and random error may have other causes related to measurement and sampling, some of which we will mention below.

Bias

The term bias refers to several statistical issues that may be classified as measurement, sampling, and estimation bias (Kotz and Johnson 1982–1988). In measurement or sampling situations, bias is "the difference between a population mean of the measurements or test results and an accepted reference or true value" (Bainbridge 1985). Therefore, bias leads to an under- or overestimate of the true value. Measurement bias is mainly due to faulty measuring devices or procedures. Therefore, measurement bias usually does not disappear with increased sampling effort, as all measurements are systematically biased away from the true value (Kotz and Johnson 1982–1988, West 1999, Debanne 2000). Sampling bias is due to unrepresentative sampling of the target population. This kind of bias does not disappear with increasing sampling effort either. For example, measuring more and more females will not give an unbiased estimate of the male population, or vice versa. Estimation bias is also called systematic error (Fig. 1a), and it refers to an estimation method for which the average of repeated estimates deviates from the true value (West 1999). Thus, estimation bias is due to the estimator itself being biased. However, estimation bias should decrease with increasing sampling effort, as this is one of the desirable characteristics of an estimator. For example, the observed number of species is a negatively biased estimator of the total species richness whose bias decreases with increasing sampling effort (Fig. 2). Below, we are only concerned with estimation bias.

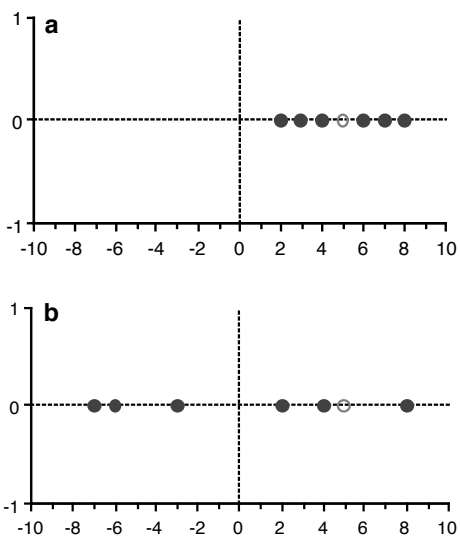

Fig. 1. Examples of (a) systematic error and (b) random error. In this hypothetical example, the true value is assumed to be zero. Note that the values on the vertical axis are presented only for graphical reasons.

Precision

Random error (Fig. 1b) is also called variability or variance, but it is also often defined as the opposite, namely precision, referring to the absence of random error. Unlike bias, its magnitude is only dependent on the estimated (or observed) values and is completely independent of the true value. Precision is thus a measure of "the statistical variance of an estimation procedure" (West 1999) or, in sampling situations, the "spread of the data ... attributable to the statistical variability present in the sample" (Debanne 2000). In measurement situations, precision arises from the variance produced by the measurement device or procedure. The total variance then arises from the variability generated by measurement error, sample variation and estimation variance. For example, the precision of measurements of a continuous variable depends on the resolution of the measuring device. The resolution of a measuring device is defined as the smallest distance over which it is possible for a value to change (Jones 1997), and the smaller this distance, the greater the variance the measuring device can detect. For example, if the resolution of the measuring device is very coarse, it will each time return exactly equal measurements of a fixed object, and precision will be zero, but if the resolution is sufficiently fine, precision will be greater than zero.

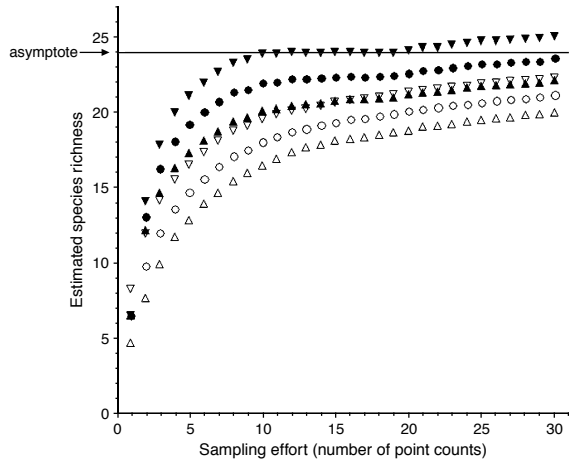

Fig. 2. Example of the bias and precision associated with two species richness estimators. The total species richness is assumed to be 24 species (example taken from Walther and Martin 2001). As sampling effort increases, the mean of the observed species richness (open circles) and the mean of the first-order jackknife estimator (closed circles) are approaching the total species richness asymptote. Standard deviations around the mean are represented by open (observed species richness) and closed (first-order jackknife estimator) triangles, respectively. Means and standard deviations were calculated from 1000 estimates derived from 1000 runs of randomized sampling order using the program EstimateS (Colwell 2000). For example, the difference between the mean of the first-order jackknife estimator and the total species richness is a measure of the bias of the first-order jackknife estimator, and the standard deviation of the first-order jackknife estimator is a measure of the precision of the first-order jackknife estimator.

The significant figures of reported measurements should reflect the precision of these measurements (Kotz and Johnson 1982–1988).

Accuracy

Bias and precision combine to define the performance of an estimator. The more biased and the less precise an estimator is, the worse its overall ability to make an accurate point estimation. Accuracy is thus defined as the overall distance between estimated (or observed) values and the true value (Bainbridge 1985, Zar 1996, Jones 1997, Krebs 1999). There are different mathematical definitions of this distance (see below), and some explicitly combine bias and precision in their mathematical definitions.

Bias, precision and accuracy, as defined above, are qualitative concepts. To quantify the performance of estimators, we now need quantitative measures that can tell us what the estimated bias, the estimated precision and the estimated accuracy of an estimator is. These estimated measures are point estimates of these qualitative concepts, using data to calculate them. For example, because the concept of accuracy incorporates the concept of bias, estimating bias and accuracy (but not precision) in a real-world situation is dependent on actually knowing (or at least guessing at) the true value of the population parameter (e.g. total species richness; see section Determining total species richness below). Below, we first introduce various bias, precision and accuracy measures (providing formulas for each of these performance measures in Table 1), and then illustrate their use in Table 2 with a worked example using some of the data presented in Fig. 2.

Unscaled performance measures

Definitions are as follows: let A be the asymptotic or total species richness (the "true value" which is a constant for any community defined in time and space, but may be different for different communities), E_j be the estimated species richness for the j^{th} sample, and n be the number of samples. In the following we assume that all estimates have been calculated for the same community, so that A is a constant. Later, we discuss how to scale performance measures so that performance measures can be used to compare performance across communities in which A is not a constant.

Bias measures

A good estimator should be unbiased, so that an even distribution of under- and overestimates leads to an overall bias of zero (Kotz and Johnson 1982–1988, Stuart and Ord 1991). Bias measures typically take into account the difference between the estimated and the total species richness.

1) One common bias measure called mean error (ME) is the mean of all differences between the estimated values and the true value (Table 1; e.g. Zelmer and Esch 1999). It indicates whether the estimator consistently under- or overestimates the total species richness. This measure has also been called mean deviation (MD) (Palmer 1990, 1991), mean difference (Rosenberg et al. 1995), mean bias (Pledger 2000) or bias (Hellmann and Fowler 1999, Foggo et al. 2003a, b).

2) Another simple bias measure is the percentage of estimates that overestimates the total species richness A (Palmer 1990, 1991, Walther and Morand 1998, Chiarucci et al. 2001, 2003, Foggo et al. 2003b, Melo et al. 2003). In this case, an unbiased estimator should return 50% overestimates and 50% underestimates.

Precision measures

A good estimator should be precise, so that its estimates show little variation (Kotz and Johnson 1982–1988, Stuart and Ord 1991). Generally, the precision of an estimator increases linearly with the square root of the sampling effort (Marriott 1990). In principle, any measure of the variability of the estimates themselves

Table 1. Performance measures of bias, precision, and accuracy. Full names for the abbreviations of performance measures are given in the text. A is the asymptotic or total species richness, E_j is the estimated species richness for the j^{th} sample, and n is the number of samples. Σ *denotes the summation formula.*

Measure	Bias	Precision	Accuracy
Unscaled	$ME = \frac{1}{n}\Sigma_{j=1}^{n}(E_j - A)$	$Var^{1)} = \frac{1}{n}\Sigma_{j=1}^{n}(E_j - \bar{E})^2$	$MSE = \frac{1}{n}\Sigma_{j=1}^{n}(E_j - A)^2$
		$SD = \sqrt{\frac{1}{n}\Sigma_{j=1}^{n}(E_j - \bar{E})^2}$	$RMSE = \sqrt{\frac{1}{n}\Sigma_{j=1}^{n}(E_j - A)^2}$
			$MAE = \frac{1}{n}\Sigma_{j=1}^{n}\lvert E_j - A\rvert$
Scaled	$SME = \frac{1}{An}\Sigma_{j=1}^{n}(E_j - A)$	$CV = 100SD/\bar{E}$	$SMSE = \frac{1}{A^2 n}\Sigma_{j=1}^{n}(E_j - A)^2$
	$PAR = \frac{1}{n}\Sigma_{j=1}^{n}(100E_j/A)$		$SRMSE = \frac{1}{A}\sqrt{\frac{1}{n}\Sigma_{j=1}^{n}(E_j - A)^2}$
	Absolute values |SME| or |PAR|		$SMAE = \frac{1}{An}\Sigma_{j=1}^{n}\lvert E_j - A\rvert$

[1)] Note that we here use the biased estimate of variance because when calculating MSE, the bias in this estimate is exactly balanced by the bias generated using the bias measure provided. If calculating variance estimates for other purposes, we recommend using the unbiased variance estimate $Var = \frac{1}{n\text{-}1}\Sigma_{j=1}^{n}(E_j - \bar{E})^2$.

can be used as a precision measure. However, typically, one of the following measures is used.

Table 2. A worked example illustrating the calculation of performance measures for two estimators (the observed species richness and the first-order jackknife estimator presented in Fig. 2). For ease of calculation, we picked only the first 10 estimates produced during the first 10 runs for each estimator at sample 30 (the highest sampling effort depicted in Fig. 2). The estimates are (presented in ascending order) 18, 19, 20, 20, 20, 20, 20, 20, 21, and 23 for observed species richness and 21.01, 21.59, 22.21, 23.53, 23.87, 23.96, 24.06, 24.69, 24.84, and 25.42 for the first-order jackknife estimator. The resulting values of the bias, precision, and accuracy measures, both unscaled and scaled (see Table 1), are presented below.

Concept	Measure	Observed species richness	First-order jackknife
Unscaled measures			
Bias	ME	−3.900	−0.482
Precision	Var	1.490	1.907
	SD	1.221	1.381
Accuracy	MSE	16.700	2.139
	RMSE	4.086	1.463
	MAE	3.900	1.084
Scaled measures			
Bias	SME	−0.163	−0.020
	PAR	83.750	97.992
Precision	CV	6.073	5.872
Accuracy	SMSE	0.029	0.004
	SRMSE	0.170	0.061
	SMAE	0.163	0.045

1) The most common precision measure is the variance (Table 1; e.g. Hellmann and Fowler 1999, Foggo et al. 2003a, b). The variance is useful as it can be combined with the mean error to measure accuracy (see below).

2) Another precision measure is the standard deviation (SD) (Table 1). For example, Baltanás' (1992) calculated the standard deviation of his bias measure PAR (see below), and Brose et al. (2003) and Melo et al. (2003) used a similar approach. Closely related to the variance, the advantage of the standard deviation is that it is on the same scale as the mean and is thus directly comparable.

3) A very simple precision measure is the range (Tietjen 1986). However, it is of essentially no use in this context because it increases with sample size (Sokal and Rohlf 1995). On the other hand, the inter-quartile or semi-interquartile range (the difference between the 25 and 75% quartile) is less sensitive to sample size and may thus be used as a more robust precision measure (Sokal and Rohlf 1995, Gould and Pollock 1997, Peterson and Slade 1998).

It is important to note that the calculation of precision measures does not require knowledge of the true value (in this case, the total species richness A). Therefore, precision measures alone cannot evaluate estimator

performance. For example, Chazdon et al. (1998) and Longino et al. (2002) suggested that an estimator should remain stable as sample size increases. However, this is just another way of suggesting to measure the precision of the estimates. Therefore, while "stability" is a quality an estimator should have with increasing sample size, it alone does not guarantee the accuracy of the estimator as it does not consider bias.

Accuracy measures

A good estimator should be accurate, so that its estimates are as close to the true value as possible (Kotz and Johnson 1982–1988, Stuart and Ord 1991). Like bias measures, accuracy measures typically take into account the difference between the estimated and the total species richness, but then square this difference or take the absolute value of it to eliminate the direction of the difference. Thus, only the magnitude of the difference is taken into account.

1) A common accuracy measure called mean square error (MSE) is the mean of the squared differences (Table 1; e.g. Burkholder 1978, Kotz and Johnson 1982–1988, Marriott 1990, Hellmann and Fowler 1999, Foggo et al. 2003a, b). It indicates how close the estimator is to the true value. This measure has also been called mean squared error (Tietjen 1986, Walsh 1997, Zelmer and Esch 1999) or mean square deviation (MSD) (Palmer 1990, 1991). This measure incorporates concepts of both bias and precision as the MSE is actually equal to the variance of the estimates plus the squared mean error (a proof of $MSE = variance + bias^2 = variance + ME^2$ can, for example, be found on p. 303 in Casella and Berger 1990). Small variance (= high precision) and little bias thus lead to a highly accurate estimator. On the other hand, an inaccurate estimator may be due to high variance and/or large bias. Since the MSE squares all differences, this measure does not have the same scale as the original measurement.

2) To return to the original scale, we can take the square root of the MSE (which is the same basic operation as turning the variance into the standard deviation). This mathematical operation yields the second accuracy measure called root mean square error (RMSE) (Table 1; e.g. Rosenberg et al. 1995, Stark 1997–2002, Zelmer and Esch 1999). Since both MSE and RMSE are calculated using squared differences, they tend to be dominated by outlying estimates far away from the true value.

3) To avoid this potential problem of outlying estimates, one may take the absolute value of the difference between the estimated and total species richness as a measure of accuracy. One can then take the mean (called mean absolute error (MAE), see Table 1; e.g. Burkholder 1978, Kotz and Johnson 1982–1988) or the median of all absolute differences (called median absolute deviation (MAD), see Norris and Pollock 1996, 1998, Pledger 2000) yielding more robust (i.e. less sensitive to outliers) measures of accuracy.

4) A simple accuracy measure is the percentage of estimates falling within the range $A \pm ([r*A]/100)$ which translates into a r% range around the total species richness asymptote A (Baltanás 1992, Walther and Morand 1998).

Scaled performance measures

One of the problems with studies evaluating estimator performance is that their results are often incomparable to other studies because the unscaled performance measures introduced above have been calculated for a specific population or community. Since these measures are not scaled according to the species richness of the community, it is invalid to compare results from communities with differing species richness. To make results comparable, all the above measures need to be scaled by dividing through the total species richness A. For example, if estimated and total species richness are 9 and 10, respectively, then the root mean square error is 1. If estimated and total species richness are 90 and 100, however, the root mean square error is 10. Although the root mean square error differs numerically, both estimates are 10% less than the total species richness, and therefore should be considered equally accurate. Therefore, scaled performance measures should be used whenever results from communities with differing species richness are compared. Later we discuss how scaled performance measures calculated from different communities may be combined to give an overall performance measure.

Bias measures

1) Scaling the ME by dividing through the total species richness A yields the scaled mean error (SME) (Table 1). This measure has also been called mean relative error (MRE) (Chiarucci et al. 2001, 2003), relative bias (Otis et al. 1978, Wagner and Wildi 2002), mean bias (Brose et al. 2003), or bias (Walther and Morand 1998, Walther and Martin 2001).

2) The SME can easily be transformed into a second scaled bias measure called percent of actual richness (PAR) (Table 1, e.g. Baltanás 1992). This measure has also been called percent of true or total richness (PTR) (Brose 2002, Herzog et al. 2002). These two measures are related as $PAR = 100*SME + 100$.

3) One can also take the absolute value of the two bias measures above to get a directionless measure of bias (e.g. Rosenberg et al. 1995). This operation is useful if only the magnitude and not the direction of the bias is of interest, but note that this is not strictly abiding to the definition of bias anymore resulting in yet another performance measure altogether.

Precision measures

1) The most commonly used measure is the coefficient of variation (CV) which provides a scaled measure of precision for values with different means (Table 1; e.g. Cam et al. 2002). It is simply the standard deviation expressed as a percentage of the mean (Sokal and Rohlf 1995).

Accuracy measures

1) Scaling the MSE by dividing through the squared total species richness A yields the scaled mean square error (SMSE) (Table 1). This measure has also been called mean inaccuracy (Brose et al. 2003), mean relative variance (Wagner and Wildi 2002), mean square relative error (MSRE) (Chiarucci et al. 2003), mean square proportional deviation (MSPD) (Palmer 1990, 1991), mean square relative deviation (MSRD) (Chiarucci et al. 2001), or deviation (Walther and Morand 1998). Unfortunately, this measure was also called precision (Walther and Martin 2001) which resulted from a somewhat confusing use of the term by Zar (1996, p. 18) who defined accuracy as "the precision of a sample statistic". Although he stated that "the precision of a sample statistic, as defined here, should not be confused with the precision of a measurement", Walther and Martin (2001) proceeded to use the term precision instead of accuracy. In the future, we recommend to use the term accuracy instead of "precision of a sample statistic" to avoid confusion.

2) Scaling the RMSE by dividing through the total species richness A yields the scaled root mean square error (SRMSE) (Table 1).

3) Scaling the MAE or the MAD by dividing through the total species richness A yields the scaled mean absolute error (SMAE) (Table 1) or the scaled median absolute deviation (SMAD).

Evaluating performance over many communities

Often we would like to evaluate an estimator's performance over a number of different communities with different species richnesses. To do this, we typically seek an average of the performance measures of interest over the different communities to produce a single value for each performance measure. However, a few points should be noted in this case.

Most importantly, scaled performance measures should always be used as unscaled measures will give higher weights to more species-rich communities (see section Scaled performance measures above). For many performance measures (e.g. SME and SMSE), the value of the measure calculated using combined data from more than one community corresponds to the weighted average of the measure calculated for each community separately (with the weights being equal to the sample size for each community). However, this is not true for all performance measures (notably variance and CV), and so care must be taken using this approach. For example, the variance of estimates derived from two separate communities may be 0.5 and 1.5, but the variance of estimates derived from the combined data may be much greater than the average of 0.5 and 1.5, depending on the difference between the means of the communities. The combined variance indicates the overall variation between the two communities rather than the average precision of the estimates within one community. This problem cannot be avoided by scaling, e.g. by using the coefficient of variation. Again, the CV of the estimates derived from combined data may be much greater than the average of the CVs derived from each separate community.

Therefore, we recommend that performance measures are calculated for each community separately and then combined as averages after the calculation. As well as avoiding the necessity of figuring out the validity of a combined data approach, calculating the measures for each community separately has a couple of additional advantages. First, it allows one to see the different contributions that each community makes towards the overall performance of an estimator, facilitating the identification of any unusual occurrences or between-community patterns in estimator performance. Second, one can choose whether a weighted average (each community estimate is weighed according to the number of samples used to calculate it) or a simple average (each community estimate is weighed equally) is more appropriate. The latter option may be reasonable if sample sizes vary widely between communities. For example, a small boreal forest will require less sampling effort than a large tropical forest to get a reliable idea of its species richness, but we may wish to consider both communities equally important when evaluating overall estimator performance.

Other performance measures

Measures of estimator performance should be viewed as suggestive, not prescriptive. Each measure should be adopted to explore an estimator's properties, but not be seen as the ultimate judgment tool, given that there is a virtually infinite number of ways of defining estimator performance in addition to the ones mentioned above. For example, another suggestion was to determine the smallest sub-sample size needed to estimate the total species richness (Melo and Froehlich 2001) while Pitman's closeness measure calculates the frequency with which one estimator is closer in absolute value to the true value than another estimator (Mood et al. 1974). Yet another suggestion is to test the ability of estimators to reliably rank communities by calculating the

regression between estimated and total species richness (Palmer 1990, 1991, Baltanás 1992, Brose 2002). However, since estimated species richness changes with sampling effort, this approach may be misleading unless estimated species richness is calculated for various levels of sampling effort (Walther and Morand 1998). Brose et al. (2003) recently used r^2-values of the regression as a precision measure, and the difference between the observed slope of the regression and the expected slope of one forced through the origin as a bias measure.

The statistical theory of point estimation has introduced several other concepts of how to evaluate estimator performance (Burkholder 1978, Kotz and Johnson 1982–1988, Lehmann 1983, Tietjen 1986, Marriott 1990, Stuart and Ord 1991). For example, a more efficient estimator has a smaller variance than another, and a consistent estimator converges on the true value as sampling effort approaches infinity. Consistency should therefore be a property of every estimator, because an estimator that yields a biased estimate even when given all possible data is surely not adequate. Such asymptotic (i.e. large-sample, or limiting) properties of an estimator do not make them irrelevant to "real-world" biological situations that usually involve small sample sizes. Whenever direct calculation of bias, precision, and accuracy is complex or infeasible, asymptotics can yield useful numerical approximations of bias, precision and accuracy for a given finite sample size. Asymptotic properties therefore do not define optimality, but rather, approximate the quantities that do define optimality. Therefore, if the sample size is large enough, good probabilistic approximations may be drawn for the given situation, and in particular, these approximations can be used to estimate bias, precision and accuracy.

Of course, point estimation of species richness is only one side of the coin, and interval estimation is also important to establish confidence intervals for a parameter. Therefore, another important criterion for comparing estimators is the coverage probability of a confidence interval which is the percentage of confidence intervals that cover the true value over many resampled data sets, i.e. the specified probability that the estimate will fall within a pre-defined lower and upper bound around the true value.

Determining total species richness

Below, we will outline how to test the performance of various species richness estimators. One prerequisite of this approach is to somehow determine the true value of total species richness so that we can calculate the bias and accuracy of each estimate returned by the estimator (a quite different approach which is not dependent on knowing the total species richness is using data-based evaluation procedures and model selection to estimate total species richness; see section Promising new research related to species richness estimation below).

To determine total species richness from real data sets, sampling must be exhaustive. Of course, asserting that a biological community or population has been sampled exhaustively is very difficult. However, as sampling effort increases, the number of singletons (i.e. single observations of a species, for details see Colwell and Coddington 1994) typically decreases once a sufficient number of species has been found (in the beginning of sampling, the number of singletons may actually increase). As long as singletons persist in the data, there is a good chance that the total species richness has not been reached. Once singletons disappear after continued sampling, one may assume that the total species richness has been reached (which, admittedly, rarely happens in most real data collections, e.g. Novotny and Basset 2000, Mao and Colwell 2005, but see Walther and Morand 1998 and Walther and Martin 2001 for such examples). The basic idea behind using singletons is that the probability of finding a new species in an additional observation is approximately the proportion of singletons remaining to be observed (see Good 1953 and Chao and Lee 1992 for details).

A closely related way of checking the data set is to plot a randomized species accumulation curve (also called a sample-based taxon re-sampling curve, see Gotelli and Colwell 2001). As long as singletons are present in the data set, this curve will be rising (Walther 1997). However, a curve with a clear horizontal asymptote indicates that the total species richness has been reached (for an illustrated example, see Walther and Martin 2001). If the total species richness cannot be ascertained in this way, one should at least state how accurate the estimate of total species richness is (e.g. by stating an interval of possible richness values; see various examples in Table 3).

If such well-sampled real datasets are not available, another approach is of course to pre-set total species richness in some kind of population model that simulates data many thousands of times, and then to compute bias, precision and accuracy measures for various estimators using these data (see various examples in Table 3). To illustrate how to use real or simulated data to test species richness estimators, we below outline some of the most important points to keep in mind when testing the performance of species richness estimators.

Testing the performance of species richness estimators

We will now use the bias, precision and accuracy measures introduced above to show how to evaluate the performance of an estimator, using a worked

Table 3. A list of studies comparing estimator performance (first column). The second column indicates the study taxon and whether real (R) or simulated (S) data were analysed. The third column gives the accuracy, bias and precision measure used in the analyses; unscaled measures: [1)]mean error, [2)]standard deviation, [3)]mean square error, [4)]root mean square error; scaled measures: [5)]scaled mean error, [6)]percent of actual richness; [7)]scaled mean square error (see Table 1); regression measures: [8)]difference between the observed slope of the regression and the expected slope of one forced through the origin, [9)]r^2-values of the regression (see Brose et al. 2003). The fourth column ranks the estimators' performance according to the respective measure with the first estimator being the best. Abbreviations of estimators are mostly taken directly from the references and refer to the following estimators: the non-parametric estimators called ACE, bootstrap (Boot), Chao1, Chao2, ICE, the first, second, third, fourth, fifth and kth-order jackknifes (Jack1, Jack2, Jack3, Jack4, Jack5, Jack-k) and the interpolated jackknife (Jack-int); the species-abundance distribution fitting methods called Preston's log-normal, Poisson's log-normal, and Cohen's truncated log-normal; the non-asymptotic curve models called log-linear (Log-Lin) "Gleason" and log-log (Log-Log) "Arrhenius" model (also called exponential and power model, respectively); and the asymptotic curve models called Beta-P, collector's curve, finite-area, Karakassis' infinite model, modified logarithmic, modified negative exponential, modified power function, negative exponential, rational function and Weibull (see Palmer 1990, Baltanás 1992, Walther and Martin 2001, Petersen and Meier 2003, Brose et al. 2003, Foggo et al. 2003b or the specific studies listed below for definitions and references). Another asymptotic curve model is the Michaelis-Menten curve model that can be fitted in four different ways (MM-least-squares, MM-Mean, MM-Monod, MM-Runs). The observed species richness (Sobs) is printed in bold because its performance is a baseline against which the performance of the other estimators can be compared. The studies using real data sets determined the total species richness in the following ways: comprehensive floristic list (Palmer 1990, 1991, Chiarucci et al. 2001, 2003); randomized species accumulation curve had reached asymptote (Walther and Morand 1998, Foggo et al. 2003b); no singletons or doubletons in the data set, therefore the randomized species accumulation curve had reached asymptote (Walther and Martin 2001); total species richness was measured as the arithmetic mean of the estimates returned from four estimators calculated using the entire data set (Brose 2002); comprehensive avian list (Herzog et al. 2002); comprehensive faunal sampling and expert knowledge (best "guesstimate") (Petersen and Meier 2003, Petersen et al. 2003). Several comparative studies (Gimaret-Carpentier et al. 1998, Keating et al. 1998, Hellmann and Fowler 1999, Melo and Froehlich 2001, Wagner and Wildi 2002, Foggo et al. 2003a, Melo et al. 2003, Rosenzweig et al. 2003, Cao et al. 2004) were not included here because most of their results were presented in figures, not tables, which made assigning ranks difficult or impossible. Note that further reviews (Brose and Martinez 2004, O'Hara 2005) were not included here because they were at the proof-stage.

Studies	Taxon	Performance measure	Estimators listed according to performance
Palmer 1990, 1991	Plants (R)	Bias[1)]	Jack2, Jack1, Log-Lin, Boot, Preston log-normal, **Sobs**, MM-Monod, Log-Log
	Plants (R)	Accuracy[3)]	Jack1, Jack2, Boot, Log-Lin, Preston log-normal, MM-Monod, **Sobs**, Log-Log
	Plants (R)	Accuracy[7)]	Jack1, Log-Lin, Jack2, Boot, Preston log-normal, **Sobs**/MM-Monod, Log-Log
Baltanás 1992	Invertebrates (S)	Bias[6)]	Jack1, Cohen's truncated log-normal, modified power function, **Sobs**
	Invertebrates (S)	Precision[2)]	Cohen's truncated log-normal, Jack1, modified power function, **Sobs**
Poulin 1998	Parasites (S)	Bias[6)]	Jack1, Chao2, Boot, **Sobs**
Walther and Morand 1998	Parasites (S)	Accuracy[7)]	Jack1, Boot, Chao1, Chao2, Jack2, MM-Mean, MM-Runs, **Sobs**
	Parasites (S)	Bias[5)]	Chao1, Chao2, Jack1, Boot, Jack2, MM-Runs, MM-Mean, **Sobs**
	Parasites (R)	Accuracy[7)]	Chao2, Jack1, Chao1, MM-Mean, Jack2, Boot, **Sobs**, MMRuns
	Parasites (R)	Bias[5)]	Chao2, Jack1, Chao1, MM-Mean, Jack2, Boot, MM-Runs, **Sobs**
Zelmer and Esch 1999	Parasites (S)	Accuracy[6)]	Boot, Jack-k, **Sobs**
	Parasites (S)	Bias[3)]	Jack-k, Boot, **Sobs**
Chiarucci et al. 2001	Plants (R)	Accuracy[7)]	Jack1, Jack2, MM-Mean, bias-corrected Chao2, Boot, **Sobs**
	Plants (R)	Bias[5)]	Jack2, Jack1, bias-corrected Chao2, MM-Mean, Boot, **Sobs**
Walther and Martin 2001	Birds (R)	Accuracy[7)]	Chao2, Chao1, Jack2, Jack1, rational function, MM-least-squares, modified power function, MM-Runs, MM-Mean, ICE, ACE, Boot, Weibull, Beta-P, **Sobs**, modified negative exponential, finite-area, negative exponential, modified logarithmic, collector's curve
	Birds (R)	Bias[5)]	Chao2, Jack1, Chao1, modified power function, Jack2, rational function, MM-least-squares, ICE, MM-Runs, ACE, MM-Mean, Boot, beta-P, Weibull, **Sobs**, modified logarithmic, modified negative exponential, finite-area, negative exponential, collector's curve
Brose 2002	Beetles (R)	Bias[6)]	Chao1, Jack2, Jack1, Boot, **Sobs**
	Beetles (R)	Precision[2)]	Chao1, **Sobs**, Boot, Jack1, Jack2
Herzog et al. 2002	Birds (S)	Bias[6)]	x-species-list method: MM-Runs, MM-Mean, Jack2, Chao2, Chao1, ICE, ACE, Jack1, Boot, **Sobs**
	Birds (R)	Bias[6)]	x-species-list method: MM-Runs, Jack2, MM-Mean, Chao2, ICE, Jack1, ACE, Chao1, Boot, **Sobs**
Brose et al. 2003	Species (S)	Bias[8)]	Jack5, Jack4, Jack-k, Jack3 = Jack-int, Chao2, Jack2, Jack1, ICE, MM-least-squares, **Sobs**, negative exponential
	Species (S)	Precision[9)]	Jack4, Jack-k, Jack3 = Jack-int, Jack2, Jack5, Chao2, ICE, Jack1, **Sobs**, negative exponential, MM-least-squares

Table 3. Continued.

Studies	Taxon	Performance measure	Estimators listed according to performance
Chiarucci et al. 2003	Plants (R)	Accuracy[7)]	Jack1, Jack2, bias-corrected Chao2, Boot, ***Sobs***
	Plants (R)	Bias[5)]	Jack1, Jack2, bias-corrected Chao2, Boot, ***Sobs***
Foggo et al. 2003b	Invertebrates	Bias[1)]	Chao2, ICE, Chao1, ACE, Boot, Jack2, Karakassis' infinite model, Jack1, ***Sobs***
Petersen et al. 2003	Diptera	Bias[6)]	Jack2, Jack1, ICE, MM-Mean, Chao2, Boot, Chao1, ACE, ***Sobs***
Petersen and Meier 2003	Diptera	Bias[6)]	Poisson log-normal, Chao1, ACE, Preston log-normal, ***Sobs***

example (Table 2). Remember that a point estimator of a parameter (e.g. total species richness) needs to be both unbiased and precise to be accurate. In Fig. 2, a typical species richness accumulation curve (or curve of observed species richness) is approaching the total species richness asymptote of 24 species as sampling effort increases. The observed species richness is inevitably a negatively biased estimator of species richness. To do better, various species richness estimators have been developed (e.g. the first-order jackknife estimator shown in Fig. 2).

These estimators try to estimate the total species richness of a defined biological community from an incomplete sample of this community. Recent reviews list numerous species richness estimators (Bunge and Fitzpatrick 1993, Colwell and Coddington 1994, Walther et al. 1995, Flather 1996, Nichols and Conroy 1996, Stanley and Burnham 1998, Boulinier et al. 1998, Chazdon et al. 1998, Keating et al. 1998, Colwell 2000, Chao 2001, 2005, Walther and Martin 2001, Hughes et al. 2001, Williams et al. 2001, Bohannan and Hughes 2003, see also reviews of species diversity estimators by Lande 1996, Mouillot and Lepêtre 1999 and Hubálek 2000), but the methodological differences between these estimators are of no concern here. All that is important in this context is that a species richness estimator will, given various data sets, yield various estimates of the total species richness. We want to test how biased, precise and accurate these estimates are, and we want to be able to compare the estimates of different estimators to evaluate their respective performances.

Note also that, in this context, it does not matter what purpose the estimator was designed for originally. For example, the estimators Chao1, Chao2, ICE and ACE were originally designed to estimate a lower bound for species richness, while the jackknife estimators were originally designed as bias reduction methods (see Table 3 for references). Each estimator has been developed to work best under the assumptions defined by a specific population model (which is the model that defines the population structure resulting from various community parameters, i.e. total species richness, species-abundance distribution, etc.), but may of course also be tested under differing population models, which is what biostatisticians or ecologists want to do. In other words, an estimator is just a function of the data, and whether it is biased, precise or accurate depends on what the researcher aims to estimate. Therefore, if a researcher decides to use a "lower-bound" or "bias-reduction" estimator to estimate total species richness, the performance of this estimator may be evaluated and compared to the performance of any other estimator. Incidentally, Chao and Tsay (1998) and Chao et al. (2005) proved that Chao1, Chao2, ICE and ACE are legitimate estimators of total species richness under some feasible population models. It is thus statistically no problem to regard these four estimators as estimators of total species richness given certain assumptions.

In some circumstances, a given estimator may have an attached statistical theory so that measures of bias, precision, and accuracy can be estimated even from a single data set (e.g. the sample mean). However, if the statistical theory for an estimator has not been developed, is not readily available, or the working assumptions do not hold, we may attempt to estimate values for bias, precision, and accuracy measures by generating large datasets and sampling them repeatedly. Many resampling schemes exist, with a large literature attached to their theory (e.g. Efron and Gong 1983, Efron and Tibshirani 1986, Manly 1997, Davison and Hinkley 1997). Resampling must be done from some empirical distribution (defined by the parameters of some population model which is itself an estimator of the true underlying distribution, e.g. the parameters of a real biological community).

In this context, it is important to note that resampling can be done with or without replacement. This crucial distinction has important implications on whether the generated data are independent of one another and are similar to real data. They are independent of one another if we resample with replacement, but generally not when we resample without replacement, except in those cases when we resample a small enough fraction of the data so that we can confidently ignore that the resampled data are dependent on one another. Moreover, over large sample sizes, the generated data are similar to real data only if we resample with replacement. Therefore, resampling without replacement complicates statistical matters considerably and should generally not be recommended. However, we cannot make a definite recommendation here, as the decision "with or without replacement" depends on the

researcher's goals defined by the study's context. This decision is not a matter of statistical principle but a question of statistical modelling, and both options are feasible. We therefore recommend that researchers refer to standard references (see above) or consult with a biostatistician on this issue.

Because any such data simulation models and resampling schemes have certain assumptions, these assumptions should be clearly described at the outset of any study. For example, researchers should be aware of the "black-box" properties of the programs they are using. One popular program is EstimateS (Colwell 2000), but most of its users did not state whether they used sampling with or without replacement. Also, EstimateS only returns summary statistics derived from the calculation of many individual data points, making the application of some formulas given in Table 1 impossible. These two important points were missed by many authors, including the senior author of this review.

Researchers should also use real datasets in addition to simulated datasets whenever possible (e.g. Walther and Morand 1998). Simulated data allow for testing of the effects of changing community parameters (i.e. total species richness, sample size, aggregation of individuals within samples) as well as the generation of bodies of data large enough for statistical analysis. However, simulated datasets may miss real patterns of community structure (Palmer 1990, 1991). Therefore, real datasets should also be tested, but only if their total species richness is known with some certainty (see section Determining total species richness above).

Moreover, researchers should compare as many different estimators as possible in their studies. Several comprehensive reviews of estimators were cited above, and several computer programs are also available (Ross 1987, Ludwig and Reynolds 1988, Rexstad and Burnham 1991, Izsák 1997, McAleece et al. 1997, Krebs 1999, White and Burnham 1999, Hines et al. 1999, Colwell 2000, Thomas 2000, Gotelli and Entsminger 2001, Turner et al. 2001, Anon. 2002). In addition, the performance of the "observed species richness" estimator should always be included in the results as a baseline against which the performance of the other estimators can be compared.

Furthermore, a mathematical expression or an unambiguous source for all estimators (e.g. Colwell 2000) and performance measures (e.g. Table 1) should always be given. Unfortunately, many studies lack such statements, and this lack of information will lead to confusion and incomparable results.

Also, the scaled performance measures resulting from the study should, whenever possible, be presented in tables and not in figures because not knowing the actual numbers makes it difficult or impossible to assign performance ranks to estimators or compare their performance across studies.

Finally, resampling should be done for various increasing levels of sampling effort, with the goal of assessing each estimator's performance with increasing sampling effort. For example, the data displayed in Fig. 2 were derived by sampling a bird community by means of 20-min point counts with fixed radius during which the number of individuals of each species was recorded (see Martin et al. 1995 for details). Each point count is considered a sample, and the entire set of point counts is considered the data set. To calculate many estimates for each level of sampling effort, the sample order is randomized many times over. For each new random combination of samples, all estimators are used to calculate estimates of total species richness. Once this has been done many times (e.g. 1000 times), the resulting 1000 estimates can then be used to calculate the bias, precision, and accuracy of each estimator at each level of sampling effort. In Fig. 2, the mean and standard deviation of the 1000 estimates for both the observed and the first-order jackknife estimator are given.

The difference between the mean of the estimates and the total species richness is then the bias of each estimator at each level of sampling effort, and the standard deviation of the estimates yields the precision of each estimator at each level of sampling effort. A table presenting the bias, precision and accuracy of each estimator at each level of sampling effort is then the most basic and complete presentation of each estimator's performance. We present a simple example in Table 2 where we calculate most of the performance measures presented in Table 1 for 10 estimates produced by the two estimators presented in Fig. 2 at constant sampling effort.

Rather than presenting all performance measures for each level of sampling in a very large table, we may want to summarize the information in such a table to get some overall measure of performance. While such an overall performance measure may not be a truly statistical property of an estimator (as the properties of an estimator change with sampling effort, and of course depend on the data), we may still want to summarize the information in such a table which is a valid procedure as long as we clearly describe how these summary performance measures are calculated. For example, the most common approach would be to average each performance measure over all levels of sampling effort to get an all-encompassing performance measure (see section Evaluating performance over many communities above).

However, to ecologists, the performance of species richness estimators may not be of interest once the observed species richness is very close to the asymptote. Therefore, such "late" samples may not be very informative when testing estimators for the practical purposes of ecological surveys. However, very "early" samples may also not be very informative simply because sampling effort is still so low (e.g. below 5 samples) that

no estimator can be expected to perform well. Therefore, Walther and Morand (1998) and Walther and Martin (2001) argued that estimator performance should be tested at "intermediate" sampling effort when observed species richness is still increasing but nowhere near the asymptote. No matter what cut-off points for sampling effort are used, levels of sampling effort used for performance evaluation should be clearly stated by researchers.

Literature review of comparative studies of species richness estimator performance

We stated above that comparative studies of estimator performance should carefully describe the details of data simulation models and resampling schemes, also use real datasets whenever possible, compare as many different estimators as possible (including the "observed species richness"), give or clearly reference mathematical expressions for all estimators and performance measures, and present results for each scaled performance measure in numerical tables with increasing levels of sampling effort.

Unfortunately, only relatively few studies have so far kept to these criteria. Many studies of estimator performance did not use both real and simulated datasets (examples in Table 3), used very few estimators (examples in Walther and Martin 2001), and presented results mostly in figures (examples in Table 3). Nevertheless, a literature review of those studies that we knew about reveals some interesting overall trends (Table 3; see also Cao et al. 2004). As expected, the observed species richness is almost always one of the worst estimators, further supporting the notion that the use of almost any estimator is preferable to the simple species count (unless sampling has been exhaustive). In most cases, non-parametric estimators (mostly the Chao and jackknife estimators) perform better than the other estimators. Even though fitting species-abundance distributions performed well in two out of three studies in which they were tested, their overall performance cannot be evaluated at present until they are included in more comparative studies (also note the problems associated with their actual application mentioned in Colwell and Coddington 1994). Curve-fitting models, on the other hand, have been extensively tested and usually perform worse than non-parametric estimators, with a few notable exceptions. The log-linear model performed quite well in Palmer's (1990, 1991) study when used with limited sample sizes, but it cannot be used to extrapolate total species richness as it has no asymptote. The modified power function and the rational function performed reasonably well in one study (Walther and Martin 2001) and perhaps deserve further consideration. Rosenzweig et al. (2003) even found that curve models far outperformed two non-parametric estimators (ICE and kth-order jackknife), but their study was based on the analysis of just one real regional data set of butterflies. Somewhat surprisingly, the Michaelis-Menten curve model performed best in one study (Herzog et al. 2002), which is contrary to its usual mediocre performance. However, Herzog et al. (2002) manipulated the data with the so-called x-species-list method prior to analysis which may explain their somewhat contradictory results. The usually superior performance of non-parametric estimators may be due to the fact that they, unlike curve models, have been developed from underlying models of detection probability (Cam et al. 2002).

To further summarize the results of our review, we calculated overall bias and overall accuracy using the information contained in Table 3. However, just as there are many different ways of evaluating bias, precision and accuracy, there are obviously various ways of summarizing the information contained in Table 3, and the particular analysis presented in Table 4 is just one of many possible ones (and we encourage researchers to do their own analysis). Nevertheless, our particular analysis further corroborates that the Chao and jackknife estimators usually perform better than the other methods, and that the observed species richness is the worst estimator. Therefore, our simple numerical analysis supports the overall qualitative impression of Table 3 presented above.

Of course, even the Chao and jackknife estimators may sometimes perform badly, and the reasons for varying performance are dependent on those variables which change the structure of the data that is used by the estimators to calculate their estimates: they are, specifically, 1) total species richness, 2) sample size, and 3) variables that change the aggregation of individuals within samples, e.g. the species-abundance distribution or the sampling protocol. In other words, while sample size and total species richness determine the actual size of the two-dimensional species-versus-sample data matrix, the species-abundance distribution and the sampling protocol determine how individuals are distributed within the individual samples, and this in return influences estimator performance. Therefore, there are no estimators that are suitable for all situations, or that are especially suitable for particular taxa, e.g. spiders or birds, unless their performance is tied to the species-abundance distribution of that taxon and the actual sampling protocol used for that taxon.

Promising new research related to species richness estimation

The development and testing of species richness estimators is an exciting and rapidly advancing field with

Table 4. Ranking estimators according to their overall bias and accuracy as summarized from the results of Table 3. For each study in Table 3, each estimator was ranked with the best estimator ranked highest. Rank was then divided by the number of estimators tested in each respective study to yield scaled ranks (resulting in 1 for the best estimator and 1/n for the worst estimator, with n being the number of estimators tested in each respective study). Scaled ranks were added over all studies and then divided by the number of studies to yield overall bias and overall accuracy. Estimators whose bias or accuracy was evaluated in <4 studies were excluded from the analysis (numbers in brackets are the number of studies in which bias or accuracy were evaluated); therefore, no results for precision are presented. Estimators are ranked here with the least overall biased estimator placed on top.

Estimator	Overall bias	Overall accuracy
Chao2	0.902 (8)	0.656 (4)
Jack2	0.826 (11)	0.645 (8)
Jack1	0.732 (14)	0.961 (7)
Chao1	0.718 (9)	– (3)
MM-Runs	0.645 (5)	– (3)
ICE	0.630 (6)	– (1)
MM-Mean	0.606 (7)	0.567 (4)
ACE	0.473 (6)	– (1)
Boot	0.437 (13)	0.601 (8)
Sobs	0.197 (15)	0.253 (15)

several publications coming out every month. It is therefore impossible to summarize all present and possible future developments, and we ask for forgiveness if we left out some promising research that we did not know about.

As mentioned above, a different approach which is not dependent on knowing the total species richness is using data-based evaluation procedures and model selection to estimate total species richness. This approach uses either goodness-of-fit or model selection criteria. Goodness-of-fit criteria (e.g. Samu and Lövei 1995, Flather 1996, Winklehner et al. 1997) essentially assume that the estimator which fits the data best will also yield the best estimate. However, the use of goodness-of-fit criteria may easily lead to over-fitted models because "increasingly better fits can often be achieved by using models with more and more parameters" (Burnham and Anderson 1998, p. 27). Therefore, other model selection criteria have been developed, of which there are many (e.g. Stanley and Burnham 1998). For example, likelihood ratio tests and discriminant function procedures have been used to choose between various jackknife estimators (e.g. Otis et al. 1978, Rexstad and Burnham 1991, Norris and Pollock 1996, Boulinier et al. 1998). However, much recent research has shown that model selection based on Kullback-Leibler information theory and maximum likelihood approaches is superior in choosing and weighting among several candidate models (Burnham and Anderson 1998, 2001, Anderson et al. 2000). These data-based model selection methods are based on the principle of parsimony "in which the data help 'select' the model to be used for inference" (Burnham and Anderson 1998, p. 27). A good example of this approach is provided in Burnham and Anderson (1998, pp. 71–72) where they re-analyse Flather's (1996) species-accumulation curve models. Model selection methods are especially useful when there is no extrapolation involved, e.g. the estimation of survival probabilities, detection probabilities, movement probabilities, and so on.

However, model selection based on maximum likelihood is only applicable if a log-likelihood function can be calculated (Burnham and Anderson 1998). Most estimators lack such a function, but so-called mixture models have recently been developed that use maximum likelihood estimation for populations with heterogeneous capture (or detection) probabilities (Norris and Pollock 1996, 1998, Gould and Pollock 1997, Gould et al. 1997, Pledger 2000, Chao et al. 2000, Pledger and Schwarz 2002). In particular, Pledger (2000) developed maximum likelihood estimators for all eight capture-recapture models developed by Otis et al. (1978) and likelihood ratio tests to choose between these models.

Besides these new non-parametric estimators, new accumulation curve models have also been developed. Following papers by Soberón and Llorente (1993), Nakamura and Peraza (1998), Keating et al. (1998), Christen and Nakamura (2000), Gorostiza and Díaz-Francés (2001), Díaz-Francés and Gorostiza (2002), Colwell et al. (2004), Mao and Colwell (2005) and Mao et al. (2005) included nonhomogeneous pure birth processes, maximum likelihood estimation and Bayesian methods into the development and comparison of curve models while Picard et al. (2004) developed a curve model that can deal with different spatial patterns.

All these new and exciting approaches to estimating species richness should be comparatively tested on real and simulated biological data. We hope that the various performance measures presented in Table 1 will help researchers to evaluate the performance of various estimators given different datasets and sampling protocols.

Acknowledgements – We thank Jean-Louis Martin for providing data, and David Anderson, Thierry Boulinier, Kenneth Burnham, Peter Caley, Douglas Clay, Robert Colwell, Paul Doherty, Curtis Flather, Gary Fowler, Rhys Green, Jessica Hellmann, Jeffrey Holman, Nils Tödtmann and Gary White for comments at various stages of the manuscript. We give very special thanks to Anne Chao, Miguel Nakamura, and Søren Feodor Nielsen who helped tremendously with very extensive comments without which we could not have written this paper.

References

Anderson, D. R., Burnham, K. P. and Thompson, W. L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. – J. Wildl. Manage. 64: 912–923.

Anon. 2002. Species diversity and richness III. Ver. 3.0, <http://www.pisces-conservation.com/indexsoftdiversity.html>.

Bainbridge, T. R. 1985. The Committee on Standards: precision and bias. – ASTM Standardization News 13: 44–46.
Baltanás, A. 1992. On the use of some methods for the estimation of species richness. – Oikos 65: 484–492.
Bohannan, B. J. M. and Hughes, J. 2003. New approaches to analyzing microbial biodiversity data. – Curr. Opinion Microbiol. 6: 282–287.
Boulinier, T. et al. 1998. Estimating species richness: the importance of heterogeneity in species detectability. – Ecology 79: 1018–1028.
Brose, U. 2002. Estimating species richness of pitfall catches by non-parametric estimators. – Pedobiologia 46: 101–107.
Brose, U. and Martinez, N. D. 2004. Estimating the richness of species with variable mobility. – Oikos 105: 292–300.
Brose, U., Martinez, N. D. and Williams, R. J. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. – Ecology 84: 2364–2377.
Bunge, J. and Fitzpatrick, M. 1993. Estimating the number of species: a review. – J. Am. Stat. Assoc. 88: 364–373.
Burkholder, D. L. 1978. Point estimation. – In: Kruskal, W. H. and Tanur, J. M. (eds), International encyclopedia of statistics. Free Press, New York, USA, pp. 251–259.
Burnham, K. P. and Anderson, D. R. 1998. Model selection and inference: a practical information-theoretic approach. – Springer.
Burnham, K. P. and Anderson, D. R. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. – Wildl. Res. 28: 111–119.
Cam, E. et al. 2002. On the estimation of species richness based on the accumulation of previously unrecorded species. – Ecography 25: 102–108.
Cao, Y., Larsen, D. P. and White, D. 2004. Estimating regional species richness using a limited number of survey units. – Ecoscience 11: 23–35.
Casella, G. and Berger, R. L. 1990. Statistical inference. – Duxburry Press, Belmont.
Chao, A. 2001. An overview of closed capture-recapture models. – J. Agricult. Biol. Environ. Stat. 6: 158–175.
Chao, A. 2005. Species richness estimation. – In: Balakrishnan, N., Read, C. B. and Vidakovic, B. (eds), Encyclopedia of statistical sciences, 2nd ed. Wiley, in press.
Chao, A. and Lee, S.-M. 1992. Estimating the number of classes via sample coverage. – J. Am. Stat. Assoc. 87: 210–217.
Chao, A. and Tsay, P. K. 1998. A sample coverage approach to multiple-system estimation with application to census undercount. – J. Am. Stat. Assoc. 93: 283–293.
Chao, A., Chu, W. and Hsu, C. H. 2000. Capture-recapture when time and behavioural response affect capture probabilities. – Biometrics 56: 427–433.
Chao, A., Shen, T.-J. and Hwang, W.-H. 2005. Application of Laplace's boundary mode approximation to estimate species and shared species richness. – Aust. N. Z. J. Stat., in press.
Chazdon, R. L. et al. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica. – In: Dallmeier, F. and Comiskey, J. A. (eds), Forest biodiversity research, monitoring and modeling: conceptual background and Old World case studies. Parthenon Publ. Group, Paris, pp. 285–309.
Chiarucci, A., Maccherini, S. and De Dominicis, V. 2001. Evaluation and monitoring of the flora in a nature reserve by estimation methods. – Biol. Conserv. 101: 305–314.
Chiarucci, A. et al. 2003. Performance of nonparametric species richness estimators in a high diversity plant community. – Div. Distrib. 9: 283–295.
Christen, J. A. and Nakamura, M. 2000. On the analysis of accumulation curves. – Biometrics 56: 748–754.
Colwell, R. K. 2000. EstimateS: statistical estimation of species richness and shared species from samples. – <http://viceroy.eeb.uconn.edu/EstimateS>.
Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity through extrapolation. – Phil. Trans. R. Soc. B 345: 101–118.
Colwell, R. K., Mao, C. X. and Chang, J. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. – Ecology 85: 2717–2727.
Davison, A. C. and Hinkley, D. V. 1997. Bootstrap methods and their application. – Cambridge Univ. Press.
Debanne, S. M. 2000. The planning of clinical studies: bias and precision. – Gastrointestinal Endoscopy 52: 821–822.
Díaz-Francés, E. and Gorostiza, L. G. 2002. Inference and model comparisons for species accumulation functions using approximating pure birth processes. – J. Agricult. Biol. Environ. Stat. 7: 335–349.
Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. – Am. Stat. 37: 36–48.
Efron, B. and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. – Stat. Sci. 1: 54–77.
Flather, C. H. 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. – J. Biogeogr. 23: 155–168.
Foggo, A. et al. 2003a. Estimating marine species richness: an evaluation of six extrapolative techniques. – Mar. Ecol. Progr. Ser. 248: 15–26.
Foggo, A. et al. 2003b. The net result: evaluating species richness extrapolation techniques for littoral pond invertebrates. – Freshwater Biol. 48: 1756–1764.
Gimaret-Carpentier, C. et al. 1998. Sampling strategies for the assessment of tree species diversity. – J. Veg. Sci. 9: 161–172.
Good, I. J. 1953. On the population frequencies of species and the estimation of population parameters. – Biometrika 40: 237–264.
Gorostiza, L. G. and Díaz-Francés, E. 2001. Species accumulation functions and pure birth processes. – Stat. Prob. Lett. 55: 221–226.
Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – Ecol. Lett. 4: 379–391.
Gotelli, N. J. and Entsminger, G. L. 2001. Ecosim: null models software for ecology, Ver. 7.0. – Acquired Intelligence and Kesey-Bear, Burlington, VT 05465, USA, <http://homepages.together.net/~gentsmin/ecosim.htm>.
Gould, W. R. and Pollock, K. H. 1997. Catch-effort maximum likelihood estimation of important population parameters. – Can. J. Fish. Aquat. Sci. 54: 890–897.
Gould, W. R., Stefanski, L. A. and Pollock, K. H. 1997. Effects of measurement error on catch-effort estimation. – Can. J. Fish. Aquat. Sci. 54: 898–906.
Hellmann, J. J. and Fowler, G. W. 1999. Bias, precision, and accuracy of four measures of species richness. – Ecol. Appl. 9: 824–834.
Herzog, S. K., Kessler, M. and Cahill, T. M. 2002. Estimating species richness of tropical bird communities from rapid assessment data. – Auk 119: 749–769.
Hines, J. E. et al. 1999. COMDYN: software to study the dynamics of animal communities using a capture-recapture approach. – Bird Study Suppl: 209–217, <http://www.mbr-pwrc.usgs.gov/software/comdyn.html>.
Hubálek, Z. 2000. Measures of species diversity in ecology: an evaluation. – Folia Zool. 49: 241–260.
Hughes, J. B. et al. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. – Appl. Environ. Microbiol. 67: 4399–4406 (erratum 2002, 68: 448).
Izsák, J. 1997. DIVERSI. Ver. 1.1. A program package to calculate diversity indices, their jackknifed estimates with confidence intervals, similarity indices and fitting abundance models. – Abstracta botanica.
Jones, C. B. 1997. Geographical information systems and computer cartography. – Longman.
Keating, K. A. et al. 1998. Estimating the effectiveness of further sampling in species inventories. – Ecol. Appl. 8: 1239–1249.
Kotz, S. and Johnson, N. L. 1982–1988. Encyclopedia of statistical sciences. Vol. 1–9. – Wiley.

Krebs, C. J. 1999. Ecological methodology. – Benjamin/Cummings.
Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. – Oikos 76: 5–13.
Lehmann, E. L. 1983. Theory of point estimation. – Wiley.
Longino, J. T., Coddington, J. and Colwell, R. K. 2002. The ant fauna of a tropical rain forest: estimating species richness three different ways. – Ecology 83: 689–702.
Ludwig, J. A. and Reynolds, J. F. 1988. Statistical ecology: a primer on methods and computing. – Wiley.
Manly, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. – Chapman and Hall.
Mao, C. X. and Colwell, R. K. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. – Ecology 86: 1143–1153.
Mao, C. X., Colwell, R. K. and Chang, J. 2005. Estimating the species accumulation curves using mixtures. – Biometrics 61: 433–441.
Marriott, F. H. C. 1990. A dictionary of statistical terms, 5th ed. – Wiley.
Martin, J. L., Gaston, A. J. and Hitier, S. 1995. The effect of island size and isolation on old growth forest habitat and bird diversity in Gwaii Haanas (Queen Charlotte Islands, Canada). – Oikos 72: 115–131.
McAleece, N. et al. 1997. BioDiversity Professional Beta Release 1. – <http://www.sams.ac.uk/dml/projects/benthic/bdpro/index.htm>.
Melo, A. S. and Froehlich, C. G. 2001. Evaluation of methods for estimating macroinvertebrate species richness using individual stones in tropical streams. – Freshwater Biol. 46: 711–721.
Melo, A. S. et al. 2003. Comparing species richness among assemblages using sample units: why not use extrapolation methods to standardize different sample sizes? – Oikos 101: 398–410.
Mood, A. M., Graybill, F. A. and Boes, D. C. 1974. Introduction to the theory of statistics. – MacGraw-Hill.
Mouillot, D. and Lepêtre, A. 1999. A comparison of species diversity estimators. – Res. Popul. Ecol. 41: 203–215.
Nakamura, M. and Peraza, F. 1998. Species accumulation for beta distributed recording probabilities. – J. Agricult. Biol. Environ. Stat. 3: 17–36.
Nichols, J. D. and Conroy, M. J. 1996. Estimation of species richness. – In: Wilson, D. E. et al. (eds), Measuring and monitoring biological diversity. Standard methods for mammals. Smithsonian Inst. Press, pp. 226–234.
Norris, J. L. and Pollock, K. H. 1996. Non-parametric MLE under two closed capture-recapture models with heterogeneity. – Biometrics 52: 639–649.
Norris, J. L. and Pollock, K. H. 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. – Environ. Ecol. Stat. 5: 391–402.
Novotny, V. and Basset, Y. 2000. Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. – Oikos 89: 564–572.
O'Hara, R. B. 2005. Species richness estimators: how many species an dance on the head of a pin? – J. Anim. Ecol. 74: 375–386.
Otis, D. L. et al. 1978. Statistical inference from capture data on closed animal populations. – Wildl. Monogr. 62: 1–135.
Palmer, M. W. 1990. The estimation of species richness by extrapolation. – Ecology 71: 1195–1198.
Palmer, M. W. 1991. Estimating species richness: the second-order jackknife reconsidered. – Ecology 72: 1512–1513.
Petersen, F. T. and Meier, R. 2003. Testing species-richness estimation methods on single-sample collection data using the Danish Diptera. – Biodiv. Conserv. 12: 667–686.
Petersen, F. T., Meier, R. and Larsen, M. N. 2003. Testing species richness estimation methods using museum label data on the Danish Asilidae. – Biodiv. Conserv. 12: 687–701.
Peterson, A. T. and Slade, N. A. 1998. Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. – Div. Distrib. 4: 95–105.
Picard, N., Karembé, M. and Birnbaum, P. 2004. Species-area curve and spatial pattern. – Ecoscience 11: 45–54.
Pledger, S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. – Biometrics 56: 434–442.
Pledger, S. and Schwarz, C. J. 2002. Modelling heterogeneity of survival in band-recovery data using mixtures. – J. Appl. Stat. 29: 315–327.
Poulin, R. 1998. Comparison of three estimators of species richness in parasite component communities. – J. Parasitol. 84: 485–490.
Rexstad, E. and Burnham, K. P. 1991. User's guide for interactive program CAPTURE. Abundance estimation of closed animal populations. – Colorado State Univ. Fort Collins, CO, <http://www.mbr-pwrc.usgs.gov/software.html>.
Rosenberg, D. K., Overton, W. S. and Anthony, R. G. 1995. Estimation of animal abundance when capture probabilities are low and heterogeneous. – J. Wildl. Manage. 59: 252–261.
Rosenzweig, M. L. et al. 2003. Estimating diversity in unsampled habitats of a biogeographical province. – Conserv. Biol. 17: 864–874.
Ross, G. J. S. 1987. Maximum likelihood program. Ver. 3.08. – Numerical Algorithm Group, Downers Grove, IL, <http://sci.agr.ca/sthyacinthe/biblio/i/n/m003032.htm>.
Samu, F. and Lövei, G. L. 1995. Species richness of a spider community (Araneae): extrapolation from simulated increasing sampling effort. – Euro. J. Entomol. 92: 633–638.
Soberón, M. J. and Llorente, B. J. 1993. The use of species accumulation functions for the prediction of species richness. – Conserv. Biol. 7: 480–488.
Sokal, R. R. and Rohlf, F. J. 1995. Biometry: the principles and practice of statistics in biological research. – Freeman.
Stanley, T. R. and Burnham, K. P. 1998. Estimator selection for closed-population capture-recapture. – J. Agricult. Biol. Environ. Stat. 3: 131–150.
Stark, P. B. 1997–2002. SticiGui: statistics tools for internet and classroom instruction with a graphical user interface. – Dept of Statistics, Univ. of California, Berkeley, USA, <http://www.stat.berkeley.edu/users/stark/SticiGui/Text/index.htm>.
Stuart, A. and Ord, J. K. 1991. Kendall's advanced theory of statistics, volume 2: classical inference and relationships – Edward Arnold.
Thomas, G. 2000. Bio-dap diversity indices. – <http://detritus.inhs.uiuc.edu/wes/populations.html>.
Tietjen, G. L. 1986. A topical dictionary of statistics – Chapman and Hall.
Turner, W., Leitner, W. and Rosenzweig, M. 2001. ws2m: software for the measurement and analysis of species diversity. – <http://eebweb.arizona.edu/diversity/>.
Wagner, H. H. and Wildi, O. 2002. Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. – Ecoscience 9: 241–250.
Walsh, S. J. 1997. Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. – Stat. Medicine 16: 669–679.
Walther, B. A. et al. 1995. Sampling effort and parasite species richness. – Parasitol. Today 11: 306–310.
Walther, B. A. 1997. Comparative studies of ectoparasite communities of birds. – Ph. D. thesis, Oxford Univ., Oxford.
Walther, B. A. and Morand, S. 1998. Comparative performance of species richness estimation methods. – Parasitology 116: 395–405.
Walther, B. A. and Martin, J. L. 2001. Species richness estimation of bird communities: how to control for sampling effort? – Ibis 143: 413–419.

West, M. J. 1999. Stereological methods for estimating the total number of neurons and synapses: issues of precision and bias. – Trends Neurosci. 22: 51–61.
White, G. C. and Burnham, K. P. 1999. Program MARK: survival estimation from populations of marked animals. – Bird Study Suppl., 46: 120–138, <http://www.cnr.colostate.edu/~gwhite/mark/mark.htm>.
Williams, B. K., Nicholls, J. D. and Conroy, M. J. 2001. Analysis and management of animal populations: modeling, estimation, and decision making. – Academic Press.
Winklehner, R., Winkler, H. and Kampichler, C. 1997. Estimating local species richness of epigeic Collembola in temperate dry grassland. – Pedobiologia 41: 154–158.
Zar, J. H. 1996. Biostatistical analysis. – Prentice Hall.
Zelmer, D. A. and Esch, G. W. 1999. Robust estimation of parasite component community richness. – J. Parasitol. 85: 592–594.

Subject Editor: Carsten Rahbek.