

# Estimation of species richness: analysis of the methods developed by Chao and Karakassis

Karl I. Ugland, John S. Gray\*

Section of Marine Biodiversity, Department of Biology, University of Oslo, Pb 1066 Blindern, 316 Oslo, Norway

**ABSTRACT:** The number of studies on marine biological diversity is increasing rapidly. One of the primary goals of such studies is to estimate how many species occur in a given area. Both parametric and non-parametric methods have been developed to make such estimates. Here, we examine 2 categories of methods: one proposed by Chao (1984) *Scand J Statist* 11:265–270 and one by Karakassis (1995) *Mar Ecol Prog Ser* 120:299–303. Foggo et al. (2003) *Mar Ecol Prog Ser* 248:15–26 recently compared these methods. By examining the bases for the methods, we show that both give large underestimates of the likely number of species in a given area, and an alternative solution is suggested.

**KEY WORDS:** Species richness · Extrapolation · Accumulation · Large samples · Sampling effort

*Resale or republication not permitted without written consent of the publisher*

## INTRODUCTION

Many countries are signatories of the Convention on Biological Diversity ([www.biodiv.org/default.aspx](http://www.biodiv.org/default.aspx)) and, as signatories, are obliged to assess the species richness within their national jurisdiction. A major part of such assessment involves estimating how many species occur in different habitats. However, quantifying species richness in practice is far more problematic than initially thought. The focus in marine research has recently addressed such issues (Karakassis 1995, Ellingsen & Gray 2002, Foggo et al. 2003) and it has become clear that (1) a proper assessment requires taxonomic expertise (especially for marine habitats which have a high phyletic diversity) and (2) most species-accumulation curves do not reach an asymptote. The problem with species identification can, in theory, be solved by using a variety of experts, and practicing quality control in sampling and sorting. In practice, the lack of taxonomic experts for the many different marine habitats such as soft-sediments is largely an economic and educational problem. The lack of an asymptote in the species-accumulation curve, however, presents a fundamentally scientific problem, which implies that all estimators have to rely on extrapolative techniques.

In a recent investigation of the performance of 6 estimators of species richness using simulation, Foggo et al. (2003) concluded that no single estimator performed the best in all cases. One of the methods studied by Foggo et al. (2003), the method developed by Chao (1984), performed best at intermediate sampling effort, while the estimator developed by Karakassis (1995) consistently underestimated. However, a modification of Karakassis's estimator (Rumohr et al. 2001) performed well at high sampling effort.

Foggo et al. (2003) used 3 assemblages: (1) beach macrofauna with an asymptote at 28 species from 59 samples; (2) estuarine oligochaetes, asymptote at 12 species from 39 samples; and (3) reef fish, asymptote of 33 species from 109 samples. Compared to many benthic communities, these 3 communities had very low numbers of species (12 to 33), which does, however, help the practical aspect of increasing the sampling effort to actually observe the asymptote of the species accumulation curve. In comparison, samples from the Norwegian continental shelf, containing 809 species, showed no sign of an asymptote even with 101 samples (Ellingsen & Gray 2002). Foggo et al. (2003) argued that the estimators should perform well at sampling efforts up to 75% of the level required to

sample all the species in the assemblage. We contend that, for most benthic assemblages, an extremely large sampling effort would need to be undertaken to fulfil this requirement. Indeed, it is unlikely that any of the existing estimators would satisfy this criterion of being an 'acceptable method' for most marine soft-sediment assemblages. Therefore, it is necessary to analyse the extent to which the methods developed by Chao (1984) and Karakassis (1995) perform when very large samples that still only collect a small fraction of the total assemblage are taken. Here, we derive these 2 families of methods in a simple probabilistic model where the underlying biological assumptions are explicitly stated. To our knowledge, such a comparison has not been performed although both methods are widely used.

## METHODS

Chao (1984) derived the estimator using highly sophisticated mathematical models that unfortunately lack any biological assumptions. The models are based on the following equation:

$$\hat{S} = S_{\text{obs}} + \frac{S_1^2}{2S_2}$$

where  $\hat{S}$  is the estimated number of species, the parameter  $S_{\text{obs}}$  represents the total number of species observed, and  $S_1$  and  $S_2$  are the number of species observed only once and twice, respectively.

Since Chao (1984) derives the estimators by reference to unusually complicated mathematical formulation (application of Riemann-Stieltjes integration and functional analysis), we derive the same estimator in Appendix 1 with techniques that only require a minimum understanding of elementary probability theory.

Karakassis (1995) suggested using a regression technique based on the idea that when sampling effort passes a certain point, no new species would be observed. He regarded sampling of new species as a catch process and, therefore, plotted the cumulative number of species present in  $k$  samples against the cumulative number of species present in  $k + 1$  samples. The method calculates the expected total number of species. When all species are caught, successive samples will contain the same cumulative number of species. Hence, a reasonable estimate of the number of species is the point where the curve intercepts the line  $y = x$ .

At first sight, this method looks quite different from the ordinary catch equations used in stock assessments of fish and marine mammals. However, this technique is in fact a member of a broad family of index methods

first developed by DeLury (1947) and later generalised to the Survival Index Method by Ugland (1985). In Appendix 2, we show that the estimate obtained with the technique developed by Karakassis (1995) is identical to the classical DeLury estimate.

In order to illustrate the difficulties of assessing species richness in a large-scaled community, we construct an assemblage with  $S_{\text{tot}} = 1000$  species containing  $N = 2558190000$  individuals according to the species abundance pattern given in Appendix 3. Following common practice, we assume a Poisson sampling and can therefore calculate all relevant quantities using a simple spreadsheet. We calculate the expected number of species not observed, and the expected number of species observed only once and twice. Each sample should contain 1000 individuals and the number of samples ( $n$ ) is assumed to increase from 1 to 1 million in the following 52 steps:

1	5	10	15	20	30
40	50	60	70	80	90
100	250	500	750	1000	2500
5000	7500	10000	15000	20000	25000
30000	35000	40000	45000	50000	55000
60000	65000	70000	75000	80000	85000
90000	95000	100000	110000	120000	130000
140000	150000	160000	170000	180000	190000
200000	400000	600000	800000	1000000	

Using these data, we analytically calculated the following quantities: (1) The expected number of observed species and Chao's (1984) estimate. Note that  $E(S_{\text{obs}})$  is the accumulation of species in this constructed assemblage and the ratio between these 2 quantities reveals how the Chao estimator behaves when applied to real data sets. (2) In order to assess the bias in Chao's estimator, we plot the relative bias (Appendix 1):

$$\text{RelBias} = E(A_0) \frac{[E(A_1)]^2}{2E(A_2)}$$

This also shows how quickly the bias converges to zero.

## RESULTS

The estimates based on the 2 sets of samples from the Norwegian continental shelf are shown in Table 1 (from Ellingsen 2001, Ellingsen & Gray 2003). As in the simulations of small communities (e.g. Foggo et al. 2003), the Chao estimator performs better than the Karakassis estimator for the large species-rich benthic communities, in that it suggests that there are more species than actually found. However, there is a con-

Table 1. The Chao and Karakassis estimates of species richness in each of the 5 regions of the Norwegian continental shelf (Ellingsen 2001, Ellingsen & Gray 2002). (a) Data are from a large sample (101 samples with 68298 ind. and 809 species sampled within 5 regions) and (b) from 6 areas in a larger sample (124 aggregated samples from 3000 single samples containing 3209254 ind. and 2186 species). Ratios of the observed to expected number of species for the Chao and Karakassis estimators are shown at the bottom of each section

(a)	Region:	1	2	3	4	5	(b) Ekofisk	Sleipner	Oseberg	Statfjord	Trøndelag	Finnmark
Samples		16	21	25	19	20	17	19	26	30	22	10
Individuals		9101	25498	17074	5201	11424	836025	877785	468531	839029	150054	37830
Species		177	307	477	297	405	932	1110	1193	1515	969	589
Chao		261	355	570	400	542	1090	1255	1303	1666	1143	731
Karakassis		165	309	505	290	388	823	1127	1170	1424	904	576
Chao		1.47	1.16	1.19	1.35	1.34	1.17	1.13	1.09	1.10	1.18	1.24
Karakassis		0.93	1.01	1.06	0.98	0.96	0.88	1.02	0.98	0.94	0.93	0.98

sistent tendency of the Karakassis estimator to underestimate the total species richness in a given area and in fact to give lower estimates of species richness to those actually observed! (Starting the regression at a more favourable x-value to the right, i.e. when the cumulative number of species shows more statistical stability gives only a slight improvement.)

At first sight, the Chao estimator seems to give very plausible results (Table 1). From the formula, it is seen that all estimates must be higher than the observed number of species. On average, the estimates were 30 % higher than the observed number of species in the large samples and 13 % higher in the extreme samples. Taking the size of the regions studied into account, it seems highly unlikely that 20 samples could contain 70 % of the total number of species. A relevant factor might be somewhere between 2 and 3 and perhaps 10 times the number of samples actually taken. For the extremely large sample, the Chao and Karakassis estimators are almost equal, and therefore must seriously underestimate the true number of species. From these analyses, we can conclude that the extrapolation techniques developed by Chao and Karakassis significantly underestimate the true species richness of benthic assemblages. Our task is to find out why these 2 methods are so biased when no asymptote is observed in the species accumulation curve (i.e. at limited sampling effort).

Fig. 1 clearly shows that the Chao estimator always lies, within a small range, above the observed number of species for 'small' sampling efforts, i.e. between  $n = 1$  and 100. Remember that  $n = 100$  represents an accumulative sample of  $2000 \times 100 = 200000$  individuals, but because we have assumed random sampling of individuals, 200000 individuals will only provide us with 15% of the total number of species. Thus, although 200000 individuals represents a large sample, it must be regarded as a small effort. Between  $n =$

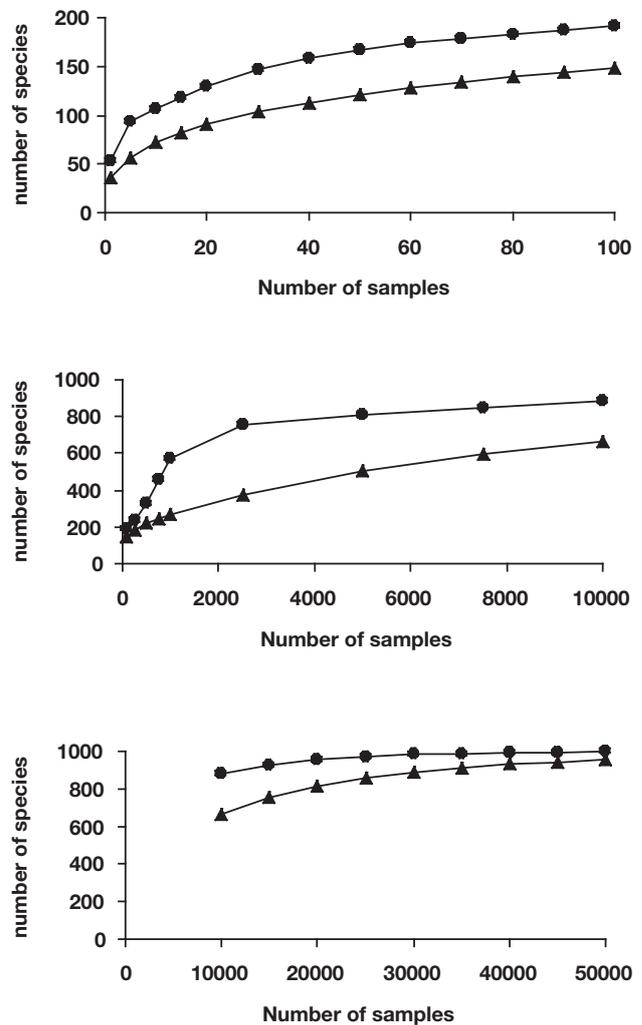


Fig. 1. Expected value of the Chao estimator (●) and the expected value of the observed number of species (▲) for increasing sample sizes in the range  $n = 1$  to 50000 samples with  $1000 \text{ ind. sample}^{-1}$

1000 and 10000 samples, the Chao estimator is considerably higher than the observed number of species. Finally, as the number of samples is increased to over 10000, the Chao estimator converges to the true number of species (1000), as does the observed number of species. Consequently, the Chao estimator is unbiased at the limit where the sampling effort becomes so great that most of the species are observed.

Fig. 2 shows why this is so. There is an extremely slow convergence towards the observed number of species; therefore, we expect that the Chao estimator also converges slowly to the true species richness. During this extremely slow convergence, Fig. 2 also shows that the ratio between the Chao estimator and the observed number of species is usually close to 1, but drifts upwards between  $n = 500$  and 5000 samples, with a peak value of 2.2 at  $n = 1000$  samples (i.e. 1 million individuals). Above  $n = 5000$ , the ratio declines monotonically to the asymptotic value of 1 that represents the required sampling effort for the Chao estimator to be unbiased.

Fig. 3 illustrates that the inherent bias in the Chao estimator is significant for our constructed assemblage. Again, the convergence to a zero bias is extremely low. At  $n = 10, 20, 100, 1000, 10000$  and 20000, the relative bias is 26.5, 23.5, 19.8, 2.4, 1.5 and 1.3, respectively.

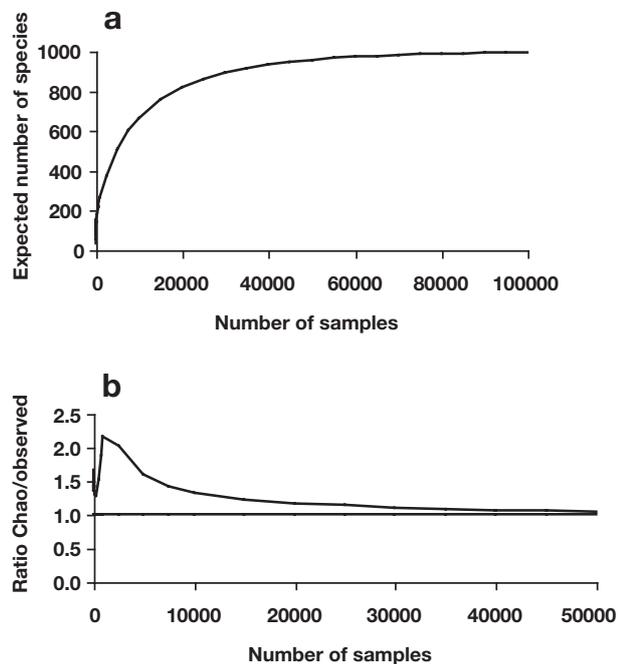


Fig. 2. (a) Accumulation curve,  $E(S_{\text{obs}})$  and (b) the ratio between the expected Chao estimator and observed number of species,  $E(S_{\text{Chao}})/E(S_{\text{obs}})$  over the whole range of number of samples ( $n = 1$  to 50000 samples)

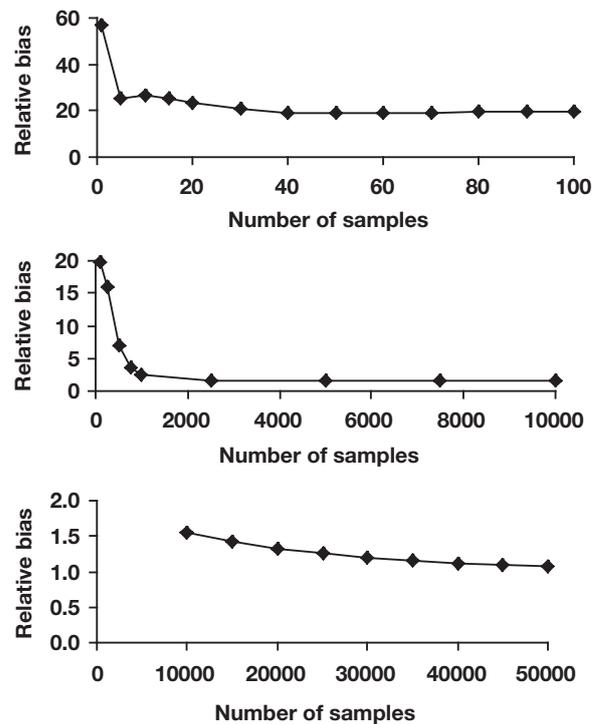


Fig. 3. Relative bias  $\text{RelBias} = E(A_0) / \frac{[E(A_1)]^2}{2E(A_2)}$  (the ratio between the expected number of species not observed and Chao's correction term; Appendix 1) calculated for sample sizes from 1 to 50000 (i.e. from 1000 to 50 million individuals)

## DISCUSSION

Foggo et al. (2003) stated explicitly that both the Chao and Karakassis estimators are effective when the habitats are quite small and discrete, with low species richness. They also conjectured that the performance of the 2 estimators is likely to differ substantially in cases with larger species richness and larger spatial scales: 'These scales will influence the frequency of encounter of rare and infrequent taxa, with deleterious consequences for the performance of the estimators'. Our results (Table 1) confirm what Foggo et al. (2003) feared would happen. Let us now try to explain this behaviour for large samples with many species.

For the Chao estimator, although it performed better than the Karakassis estimator, the improvement was relatively slight and also here the estimates were severe underestimates of total species richness. It is shown in Appendix 1 that the Chao estimator has an inherent bias:

$$\text{Bias} = E(A_0) - \frac{[E(A_1)]^2}{2E(A_2)}$$

where  $A_0$  is the number of species not observed, and  $A_1$  and  $A_2$  are the number of species observed once

and twice. This bias is the source of the difficulties in the model proposed by Chao since the mathematical statement that the bias is only an approximation (or equivalently: converges to zero with increasing sample size) is in fact a biological proposition about the relationship between unobserved species ( $A_0$ ) and rare species ( $A_1$  and  $A_2$ ).

As previously mentioned, it is doubtful whether it is possible to postulate a relationship between the species composition in future samples since they are likely to vary continuously over the investigated habitat. The assumption of a relationship between unobserved species and rarely observed species is, therefore, not likely to be satisfied since the abundance of rare species is mainly determined among the species observed, and the unobserved species are part of a covariance matrix that still has to be observed (which in all probability will be different from that obtained from the observed samples). Consequently, the Chao estimator also requires a large effort and should not be used in a rapid assessment of species richness.

The Karakassis estimator, shown in Appendix 2, is a rather old method, since it is a variant of DeLury's method (DeLury 1947). The fundamental underlying assumptions of this model are (1) that the probability of being caught for any individual is proportional to the sampling effort and (2) that the proportionality factor, called the catchability or catch coefficient, is independent of the effort and the same for all individuals. In statistical terms, this may be stated more concisely as an assumption that the catch process is a Poisson process with regard to effort. For large communities, this assumption is erroneous since the encounter of rare species is a different process than the encounter of the more common species; therefore, catchability cannot be the same for all species. It is likely that catchability decreases as the accumulated number of sampled species increases. This explains why the Karakassis estimator consistently underestimates and why regressing  $S_k$  against  $S_{k+10}$  improved the performance of plotting  $S_k$  against  $S_{k+1}$ . In light of DeLury's regression, this is simply an ad hoc technique to incorporate the concavity of the catch curve. From a statistical point of view, it would have been better to try to relax the requirement of a constant catchability by, for example, allowing it to be a decreasing function. Seber (1973) gives a short discussion of models where catchability declines exponentially.

The reason why we do not pursue the idea of generalizing DeLury's (1947) method is that we have good reason to believe that as samples accumulate in benthic assemblages, the covariances between species abundances also vary and the fraction of rare species frequently increases rather than decreases. Consequently, the species curve keeps on deviating from the

$x$ -axis (representing the sampling effort) and no asymptote is observed.

This explains why the Karakassis estimator will not function properly before a significant share of the total number of species has been observed. The trouble is, however, that we have no method to decide what constitutes 'a sufficiently large effort'; thus, the Karakassis estimator shares the common property of all members of DeLury-type estimators: they only function when a large effort has been used. Therefore, methods of the DeLury family should not be used to estimate species richness.

Traditional methods of estimating species richness (the Chao and Karakassis methods, and extrapolations from species accumulation curves such as Colwell's EstimateS programme, Colwell 2000; EstimateS: statistical estimation of species richness and shared species from samples. Version 5. User's Guide and application available at: <http://viceroy.eeb.uconn.edu/estimates>) use methods based on the randomisation of all samples collected. Any heterogeneity in the habitat, where some areas contain many species and others few species, is not taken into account. Likewise, the traditional methods assume that all species are able to colonise the whole area sampled. For the Norwegian continental shelf, this is not the case and most species are confined to narrow areas of the whole shelf (Ellingsen 2001, Ellingsen & Gray 2002). Based on these data, Ugland et al. (2003) presented an approach which gave more realistic estimates of the likely total species richness expected to be found on the Norwegian continental shelf. The method is based on recognising that there is environmental heterogeneity between sub-areas sampled and that many rare species are confined to sub-areas. The method derives a total species curve based on accumulation between sub-areas. Although the model was originally based on assuming 5 sub-areas, the model has been generalised so that any number of sub-areas can be analysed. Algorithms can be found at <http://folk.uio.no/johnsg/>. For example, for the data in Table 1a, Chao estimators predicted ca. 1100 species from the 809 species found. Yet, we have surveyed much larger areas of the shelf than that used in the data shown in Table 1a and in fact over 2500 species have been found. Thus Chao's estimator still gave a gross underestimate. The Ugland et al. (2003) method estimates ca. 5200 species, which at just over twice the number found is a reasonable number. Whether or not this is a correct estimate can only be judged after many more years of sampling. However, using the Ugland et al. (2003) method on data from surveys of the bird species in Panama, N. O'Dea & R. Whitaker (unpubl. data) predicted 428 species, whereas 427 were found; thus we are hopeful that this method may give more accurate estimates of species richness.

### Appendix 1. Derivation of Chao's estimator

Consider a community with  $S_{\text{tot}}$  species and  $N$  individuals. Denote the population size of the  $i$ th species by  $N_i$  and let  $p_i = N_i/N$  be the fraction of the  $i$ th species in the community. Now we need to aggregate all the  $p$ -values into a random variable, so we define  $U$  as the  $p$ -value of a randomly selected species among all the  $S_{\text{tot}}$  species (i.e. each species has the same chance  $1/S_{\text{tot}}$  of being selected). It is emphasized that here we draw randomly from the set  $\{1, 2, 3, \dots, S_{\text{tot}}\}$  rather than among the set of all individuals  $\{1, 2, 3, \dots, N\}$ . Assume further that the random variable  $U$  has the probability density  $f(p)$ ; that is  $P(p < U < p + \Delta p) = f(p)\Delta p$ .

Suppose we now take a sample of  $N$  individuals and define the  $S_{\text{tot}}$  random variables  $X_i =$  number of the  $i$ th species in the sample;  $i = 1, 2, \dots, S_{\text{tot}}$ . Chao follows Fischer et al. (1943) and assumes  $X_i$  has a Poisson distribution with parameter  $Np_i$ . Since  $\int_0^1 S_{\text{tot}} f(p) dp$  is the number of species that are observed  $k$  times in the samples;  $k = 0, 1, 2$ . Since the number of species that are represented between the fraction  $p$  and  $p + \Delta p$  among the individuals in the community is  $S_{\text{tot}} f(p) \Delta p$ , and all of these have the Poisson probability  $[(Np)^k/k!]e^{-Np}$  of being represented with  $k$  individuals, the expected value of  $A_k^s$  will be:

$$E(A_k) = \int_0^1 \frac{Np^k}{k!} e^{-Np} S_{\text{tot}} f(p) dp$$

The clue is now to focus on the species that have been sampled only once and define the (conditionally) random variable  $Z$ , the  $p$ -value of a species that is sampled only once. Then:

$$P(Z = p) = P(U = p | X = 1) = \frac{P(U = p \cap X = 1)}{P(X = 1)} = \frac{P(X = 1 | U = p)P(U = p)}{\sum_p P(U = p \cap X = 1)P(U = p)}$$

so the probability density of the variable  $Z$  must be given by:

$$g(p) dp = \frac{Np e^{-Np} f(p) dp}{\int_0^1 Np e^{-Np} f(p) dp}$$

In order to relate the number of unobserved species ( $A_0$ ) to the number of species observed once and twice ( $A_1$  and  $A_2$ ), we calculate  $E(Z)$  and  $E(1/Z)$ :

$$E(Z) = \int_0^1 p g(p) dp = \frac{\int_0^1 Np^2 e^{-Np} f(p) dp}{\int_0^1 Np e^{-Np} f(p) dp} = \frac{2 \int_0^1 \frac{(Np)^2}{2} e^{-Np} f(p) dp}{E(A_1)} = \frac{2E(A_2)}{NE(A_1)}$$

$$E\left(\frac{1}{Z}\right) = \int_0^1 \frac{1}{p} g(p) dp = \frac{\int_0^1 \frac{Ne^{-Np} f(p) dp}{p}}{\int_0^1 Np e^{-Np} f(p) dp} = \frac{N \int_0^1 e^{-Np} f(p) dp}{E(A_1)} = \frac{NE(A_0)}{E(A_1)}$$

It turns out that the essential part of the derivation of Chao's estimator is a direct substitution into the inequality  $E(1/Z) > 1/E(Z)$  for positive random variables that all have values  $< 1$ :

$$E\left(\frac{1}{Z}\right) = \frac{NE(A_0)}{E(A_1)} \geq \frac{1}{E(Z)} = \frac{NE(A_1)}{2E(A_2)}$$

from which we obtain the inequality:

$$E(A_0) \geq \frac{[E(A_1)]^2}{2E(A_2)}$$

Applying advanced mathematics (rather than the simple inequality  $E[1/Z] > 1/E[Z]$ ), Chao concluded (without any biological interpretation) that the best estimate for the number of unobserved species is  $A_1^2/(2A_2)$ , so the total number of species should be estimated as:

$$\hat{S}_{\text{tot}} = S_{\text{obs}} + \frac{A_1^2}{2A_2}$$

As seen here, this estimator rests heavily on 1 inequality whose accuracy is determined by the distribution of rare species. Thus, in communities where the overwhelming species are rare, the Chao estimator must be expected to have a significant large bias:

$$\text{Bias} = E(A_0) - \frac{[E(A_1)]^2}{2E(A_2)}$$

In order to make the measurement of the bias independent of the particular values of the simulation, we present the ratio which measures the relative bias:

$$\text{RelBias} = E(A_0) / \frac{[E(A_1)]^2}{2E(A_2)} = \left(\text{Bias} / \frac{[E(A_1)]^2}{2E(A_2)}\right) - 1$$

It remains to prove the fundamental inequality  $E(1/Z) > 1/E(Z)$  for random variables taking values between 0 and 1. To this end, we simplify the proof by discretizing; so if  $z$  takes the  $m$  values  $z_1$  ( $^{\circ}K$ ),  $z_m$  with probabilities  $p_1$  ( $^{\circ}K$ ),  $p_m$  the discrete form of the inequality becomes:

$$\frac{1}{z_1} p_1 + \dots + \frac{1}{z_m} p_m \geq \frac{1}{z_1 p_1 + \dots + z_m p_m}$$

which is equivalent to:

$$p_1^2 + \dots + p_m^2 + \sum_{i < j} \left(\frac{z_i + z_j}{z_j} + \frac{z_j}{z_i}\right) p_i p_j \geq 1$$

In this expression:

$$\frac{z_i + z_j}{z_j} + \frac{z_j}{z_i} = \frac{z_i}{z_j} + \frac{1}{\frac{z_i}{z_j}} = t + \frac{1}{t} \geq 2$$

where we have put  $Z_i/Z_j = t$  and used the fact that whatever the value of  $t$ ,  $(t-1)^2 = t^2 - 2t + 1 \geq 0$ , so dividing by  $t$  and rearranging gives  $t + 1/t \geq 2$ . Hence:

$$p_1^2 + \dots + p_m^2 + \sum_{i < j} \left(\frac{z_i}{z_j} + \frac{z_j}{z_i}\right) p_i p_j \geq p_1^2 + \dots + p_m^2 + 2 \sum_{i < j} p_i p_j = (p_1 + \dots + p_m)^2 \geq 1$$

which demonstrates the truth of the fundamental inequality.

**Appendix 2.** Equivalence of De Lury's and Karakassis's method

Following the ordinary underlying assumptions in the theory of Poisson processes (Feller 1970), we assume that each species has the same probability  $q\Delta f$  of being sampled when the effort  $\Delta f$  has been employed. Let  $S_i$  be the number of species included in the first  $i$  samples, so the number of new species (i.e. catch) in the  $i$ th sample will be  $C_i = S_i - S_{i-1}$ . The conditional expected catch of new species is given by De Lury's catch equation:

$$E(C_i|S_{i-1}) = q(S_{\text{tot}} - S_{i-1})$$

where  $S_{\text{tot}}$  is the total number of species. Since the line  $y_{\text{DeLury}} = q(S_{\text{tot}} - x)$  intercepts the  $x$ -axis at  $S_{\text{tot}}$ , DeLury's estimate of the total number of species is the intercept of the regression line with the  $x$ -axis. It represents the point where

the expected catch is zero. The only difference is that here we 'catch' (i.e. collect) species, while De Lury (1947) considered records of lobster catches, i.e. individuals of the same species.

To see how the Karakassis estimator is simply a reformulation of De Lury's method, substitute  $C_i = S_i - S_{i-1}$  and obtain  $E(S_i|S_{i-1}) = qS_{\text{tot}} + (1 - q)S_{i-1}$ . Sooner or later the samples will not contain new species because all species are sampled (i.e. when  $S_i = S_{i-1} = S_{\text{tot}}$ ). Since the line  $y_{\text{Karakassis}} = qS_{\text{tot}} + (1 - q)x$  intercepts the line  $y = x$  at  $S_{\text{tot}}$ , Karakassis estimated the total number of species to be the intercept of the regression line with the identity line. It remains to observe that  $E(S_i|S_{i-1}) = S_{i-1}$  is equivalent to  $E(C_i|S_{i-1}) = 0$ , which fulfills the proof that the Karakassis estimator is equivalent to DeLury's estimator.

**Appendix 3.** The model community. The parameters needed to construct the model community. The first and second columns give the population sizes and the number of species having these abundances. The number of individuals is obtained by multiplying these 2 columns, and the p-values are calculated by dividing the first column by the sum of the third column. The density  $f(p)$  is the number of species (column 2) divided by 1000 (i.e. species richness = sum of column 2)

Population size	Number of species	Number of individuals	Populations as fraction of total no. of individuals (p)	Fraction of species (among the species) with given population size (f(p))
100	300	30000	0.000000039	0.300
200	150	30000	0.000000078	0.150
300	100	30000	0.000000117	0.100
400	75	30000	0.000000156	0.075
500	50	25000	0.000000195	0.050
600	45	27000	0.000000235	0.045
700	40	28000	0.000000274	0.040
800	35	28000	0.000000313	0.035
900	30	27000	0.000000352	0.030
10000	25	250000	0.000003909	0.025
20000	20	400000	0.000007818	0.020
30000	17	510000	0.000011727	0.017
40000	15	600000	0.000015636	0.015
50000	12	600000	0.000019545	0.012
75000	11	825000	0.000029318	0.011
100000	10	1000000	0.000039090	0.010
200000	9	1800000	0.000078180	0.009
300000	8	2400000	0.000117270	0.008
400000	7	2800000	0.000156361	0.007
500000	6	3000000	0.000195451	0.006
750000	5	3750000	0.000293176	0.005
10000000	4	40000000	0.003909014	0.004
15000000	4	60000000	0.005863521	0.004
20000000	3	60000000	0.007818028	0.003
25000000	3	75000000	0.009772534	0.003
30000000	3	90000000	0.011727041	0.003
35000000	2	70000000	0.013681548	0.002
40000000	2	80000000	0.015636055	0.002
45000000	2	90000000	0.017590562	0.002
50000000	2	100000000	0.019545069	0.002
75000000	1	75000000	0.029317603	0.001
100000000	1	100000000	0.039090138	0.001
200000000	1	200000000	0.078180276	0.001
500000000	1	500000000	0.195450690	0.001
1000000000	1	1000000000	0.390901379	0.001

## LITERATURE CITED

- Chao A (1984) Non-parametric estimation of the number of classes in a population. *Scand J Statist* 11: 265–270
- DeLury DB (1947) On the estimation of biological populations. *Biometrics* 3:145–167
- Ellingsen KE (2001) Biodiversity of a continental shelf soft-sediment macrobenthic community. *Mar Ecol Prog Ser* 218: 1–15
- Ellingsen KE, Gray JS (2002) Spatial patterns of benthic diversity: Is there a latitudinal gradient along the Norwegian continental shelf? *J Anim Ecol* 71: 373–389
- Feller W (1970) *An introduction to probability theory and its applications*, 3rd edn. Wiley, New York
- Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J Anim Ecol* 12:42–58
- Foggo A, Attrill MJ, Frost MT, Rowden AA (2003) Estimating marine species richness: an evaluation of six extrapolative techniques. *Mar Ecol Prog Ser* 248:15–26
- Karakassis I (1995)  $S_{\infty}$ : a new method for calculating macrobenthic species richness. *Mar Ecol Prog Ser* 120:299–304
- Rumohr H, Karakassis I, Jensen JN (2001) Estimating species richness, abundance and diversity with 70 macrobenthic replicates in the Western Baltic Sea. *Mar Ecol Prog Ser* 214:103–110
- Seber GAF (1973) *The estimation of animal abundance*. Hafner Press, New York
- Ugland KI (1985) Abundance estimation of the Northwestern Atlantic harp seal population. *Polar Biol* 4:187–198
- Ugland KI, Gray JS, Ellingsen KE (2003) The species accumulation curve and estimation of species richness. *J Anim Ecol* 72:888–897

*Editorial responsibility: Otto Kinne (Editor), Oldendorf/Luhe, Germany*

*Submitted: March 10, 2004; Accepted: August 26, 2004  
Proofs received from author(s): December 1, 2004*