Sampling Properties of a Family of Diversity Measures

WOOLLCOTT SMITH and J. FREDERICK GRASSLE

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, U.S.A.

Summary

A family of species diversity measures proposed by Hurlbert (1971) is defined as the expected number of species in a random sample of m *individuals from a population. For* m = 2 *this measure is equivalent to Simpson's diversity index. For larger* m, *the measure is increasingly sensitive to rare species. In this paper we use unbiased estimation theory to obtain a minimum variance unbiased estimator for this family of diversity measures. An unbiased estimator of the sampling variance is also obtained. These results are then used to partition the variation in sample diversity between random sampling error and local variation community diversity.*

1. Introduction

Although species diversity is a property of a population, diversity measures are estimated, in practice, using samples from the large unknown population. Many proposed diversity indices, while theoretically appealing, have unsatisfactory sampling properties, especially when the sample size is small. Bowman, Hutcheson, Odum and Shenton (1969) have extensively reviewed the small sample properties of the commonly used Shannon-Wiener information statistic. They show that on an average the information estimated from the sample will underestimate the true population information. That is, the estimate of Shannon-Wiener information is biased; the bias is dependent on sample size with the bias going to 0 as the sample size becomes large. Unfortunately there is no simple relationship between bias and sample size since the bias will also depend on the true population diversity. For sample sizes less than 50, there is a significant underestimate of the true population diversity. Differences in diversity estimated from samples from two different populations can result from three distinct effects: true differences in population diversity, sample size dependent bias of the diversity estimator, and random sampling error. It is difficult if not impossible to separate these three effects, particularly when the sample size is small and the bias is large.

These difficulties with the small sample properties of Shannon's information have led Eberhardt (1971) and others to suggest Simpson's diversity measure for small samples. Good (1953) has given an unbiased estimate of Simpson's diversity, and its sampling variance is known. The one major criticism of Simpson's diversity is a biological one: it tends to be heavily dependent on the dominant species in the population (Williams 1964, Sanders 1968 and Whittaker 1972). A useful way of looking at any diversity index is to find each species' contribution to the total diversity index (Peet 1974). For Simpson's index each species' contribution is given by the probability that it will appear in a random sample of size two from the population (see Section 2). Thus, rare species will contribute little to the diversity index.

Key Words: Species diversity; Expected species index; Shannon information; Simpson's index; Unbiased estimation.

The reader will be delighted to find that we do not propose yet another measure of species diversity. We show that Simpson's diversity can be generalized in such a way that the good small sample properties are retained while weight is given to the rare species. The family of generalized Simpson's measures we investigate was first proposed by Hurlbert (1971), as an improvement on Sanders' (1968) rarefaction method. Hurlbert's family of diversity measures, which we call expected species diversity, is defined as the expected number of species encountered when m individuals are drawn at random from the population. We will call this a family of diversity measures since each m defines a separate diversity measure. For fixed m, each species' contribution to the diversity measure is the probability that it will appear among m individuals drawn at random from the population. For large m, the measure is sensitive to the rare species in the population while for small m, the measure is dominated by the abundant species. For $m = 2$, the measure is equivalent to Simpson's diversity.

A source of confusion surrounding Hurlbert's expected species measure results from the use of the phrase "sample size" in two different contexts. In common usage sample size refers to the number of individuals in an actual field sample from the population. This is the meaning we will use throughout this paper. Another meaning of sample size is implied in defining expected species diversity. The contribution of each species to the overall diversity index is the probability that it will occur in a random sample of size m. We will call m the individual index to clarify the distinction between the hypothetical sample of size m and the size of the actual field sample from which the diversity measure is calculated.

Expected species diversity has been called a species richness measure by Peet (1974) and Hurlbert (1971), and a diversity measure by Sanders (1968). This semantic difficulty results in part from confusion between the actual field sample size and the individual index. For fixed individual index m, the expected species measure satisfies Pielou's (1969) description of a diversity measure in that it is dependent on the number of species in the field sample and on the evenness with which the individuals in the field sample are distributed among species. Consistent with the previous discussion then, expected species measures are best described as a family of diversity measures with varying rare-species sensitivities depending on the individual index m.

A number of different justifications and interpretations have been given for expected species diversity (Sanders 1968, Hurlbert 1971, Fager 1972, Peet 1974 and Raup 1975). The range of interpretation has resulted in considerable confusion. In this paper the justification is simple and straightforward: expected species diversity generalizes Simpson's diversity to give greater rare-species sensitivity while retaining the good sampling properties of Simpson's measure. For large m, expected species diversity appears to be the only measure in the literature that is both sensitive to rare species and unbiased.

Ecologists also will have noted a more obvious semantic difficulty: we have used the general statistical meaning of the word "population," in that samples are drawn from an unknown population. In this particular application the population is the unknown multi-species community whose diversity we want to estimate. The usage in this paper is thus in unavoidable conflict with ecological usage where the word "population" usually denotes a set of organisms belonging to the same species.

In this paper, we first present a more precise definition of Hurlbert's expected species measure, show that it has an unbiased estimator, and develop an unbiased estimator of its sampling variance. In the remaining sections we give a method for investigating diversity in nonhomogeneous populations and apply the expected species diversity index to samples from natural communities.

2. Expected Number of Species

Hurlbert (1971) introduced the idea of using the expected number of species in a sample of m individuals as a measure of species diversity. Changing Hurlbert's notation slightly, suppose we have a finite population consisting of k species with n_i individuals of species i. Let the vector, $\mathbf{n} = (n_1, n_2, \cdots, n_k)$, represent the entire finite population. Let the random variable S denote the number of species in a sample of m individuals. The expected number of species in a random sample without replacement, given the finite population, $\mathbf{n}$, is

$$E[S|\mathbf{n}] = \sum_{i=1}^{k} 1 - C(n - n_i, m)/C(n, m) \tag{1}$$

where n is the total number of individuals in the finite population, $n = \sum_i n_i$. We propose the following straightforward extension in Hurlbert's idea. Suppose we have a multinomial population where π_i is the proportion of individuals of species i in the population and $\pi = (\pi_1, \pi_2, \cdots, \pi_k)$ is the vector of probabilities describing the population. The multinomial sampling model assumes the population is infinite and the identification of each individual sampled is independent of other individuals sampled. In other words, there is no patchiness within species. Under this regime the expected number of species in a sample with m individuals is

$$s(m) = E[S|\pi] = \sum_{i=1}^{k} [1 - (1 - \pi_i)^m]. \tag{2}$$

When the individual index m is two, $s(m)$ is related to Simpson's diversity, $SI = \sum_i \pi^2_i$ since $s(2) = 2 - \sum_i \pi^2_i$. For small index values, m, $s(m)$ emphasizes the more abundant species while for larger values of m the measure includes the less abundant species. For large m, of course, $s(m)$ approaches the total number of species present in the multinomial population.

A major practical advantage that Hurlbert's $s(m)$ and Simpson's index hold over other diversity measures is that both have minimum variance unbiased estimators. Suppose we have a sample of N individuals from the multinomial population π, with N_i individuals of species i. Let $\mathbf{N} = (N_1, N_2, \cdots, N_k)$ denote the vector describing the random sample. From the random sample $\mathbf{N}$ we wish to estimate $s(m)$, the expected number of species in a sample of m individuals from the population. The minimum variance unbiased estimate is

$$\hat{s}(m) = E[S|\mathbf{N}], \tag{3}$$

where $E[S|\mathbf{N}]$ is the expected number of species in a finite population given in equation (1). This result is easily obtained by a straightforward application of the Rao-Blackwell theorem (Fraser 1958, page 220). The basic result we need is that if $\hat{\theta}$ is an unbiased estimator of θ, then the minimum variance unbiased estimator is obtained by finding the expected value of $\hat{\theta}$ conditioned on a complete sufficient statistic for the sample (Rao 1965, page 26). In order to apply this result one must first find an unbiased estimator of θ even if it is a trivial estimator. For example, a single observation of a random variable is an unbiased estimator of the mean of that random variable. In our case, the random variable S, the number of species in the first m individuals drawn from the population, is a trivial unbiased estimator of $s(m)$. A complete sufficient statistic for a sample from the multinomial family is $\mathbf{N}$; thus (3) is the minimum variance unbiased (MVU) estimator of the expected number of species in a random sample from a multinomial population.

We cannot apply this theory to other diversity measures such as Shannon's information index simply because we cannot do the first step: find a trivial unbiased estimator. In fact, Blyth (1959) shows that an unbiased estimator of Shannon's information does not exist.

One can now exploit this MVU property to obtain an estimator of the variance of $\hat{s}(m)$. We note that the variance of S can be partitioned in the following way:

$$\text{Var}\,[S] = E[\text{Var}\,[S|\mathbf{N}]] + \text{Var}\,[E[S|\mathbf{N}]]. \tag{4}$$

The last term in (4) is Var $[\hat{s}(m)]$. To find an unbiased estimator for the variance of $\hat{s}(m)$, one needs unbiased estimators for the other two terms in the equation. Suppose $N \geq 2m$ and let S' and S'' denote the number of species in two random non-overlapping subsamples of size m from the finite sample $\mathbf{N}$. The two non-overlapping subsamples of size m are equivalent to two independent samples from the infinite multinomial population π. Thus $(S'' - S')^2/2$ is an unbiased estimator of their variance when sampled from a multinomial population. The minimum variance unbiased (MVU) estimator by the Rao-Blackwell Theorem is then

$$\widehat{\text{Var}}\,[S] = 1/2\, E[(S'' - S')^2|\mathbf{N}]. \tag{5}$$

One should note that S'' and S' are not independent once they are conditioned on the observed sample $\mathbf{N}$. Equation (5) is the estimator of the variance of S when m individuals are sampled from an infinite multinomial population; it is not the variance of S when m individuals are subsampled from a known finite sample, $\mathbf{N}$. The MVU estimator for the first term on the right hand side of (4) is Var $[S|\mathbf{N}]$. The estimator for Var $[\hat{s}(m)]$ is obtained by substituting these unbiased estimators into (4) and rearranging terms

$$\begin{aligned}\widehat{\text{Var}}\,[\hat{s}(m)] &= 1/2\, E[(S' - S'')^2|\mathbf{N}] - \text{Var}\,[S|\mathbf{N}] \\ &= -\text{Cov}\,[S', S''|\mathbf{N}] \\ &= -E[S' \cdot S''|\mathbf{N}] + \hat{s}(m)^2.\end{aligned} \tag{6}$$

Since this unbiased estimator is a function of a complete sufficient statistic for the multinomial family, it is therefore minimum variance unbiased. Raup (1975) has given a simple computational form for Var $[S|\mathbf{N}]$. The estimate of the sampling variance, (6), is more difficult to compute. This problem is discussed in the appendix. Approximate confidence intervals for $\hat{s}(m)$ can be found using this estimator of the variance.

3. Diversity in Nonhomogeneous Populations

The unbiased property of the expected species estimator allows us to extend this diversity estimator in a natural way to a class of nonhomogeneous populations. Since populations usually do not have homogeneous distributions in space or time, a series of samples from a particular area is usually not drawn from a single multinomial population (Fager 1972). A set of samples from a nonhomogeneous population contains information not only on the diversity but also on the spatial variability within the community. We can extend the ideas in the previous paragraphs to include the case where samples are drawn from a series of unknown multinomial subpopulations, $\pi_1, \pi_2 \cdots, \pi_J$. The probability that the ith independent random sample $\mathbf{N}_i$ is from the jth multinomial population, π_j, is $p(j)$. In words, the distributions of individuals within a sample are multinomial; however, each sample is a random sample drawn from an independently selected but unknown subpopulation.

The sampling model described is analogous to the sampling regime if, for example, a series of benthic cores were used to characterize an area. Here, the bottom community would vary with distance between the cores; however, within the core the population might reasonably be considered randomly distributed (Jumars 1975a,b). Each core then can be considered a sample from a randomly selected subpopulation of the total community.

The first problem is to define diversity in this situation. One approach would be to form an artificial pooled multinomial population and then define a "large scale" diversity measure based on the pooled population (Hessler and Jumars 1974, page 203). However, here we will define a "small scale" diversity measure $\bar{s}(m)$ as the average number of species in a sample of m individuals from the population,

$$\bar{s}(m) = \sum_{j=1}^{J} E[S|\pi_j]p(j) \ .$$

This seems to be the natural measure of diversity in the context of the expected species diversity measure. This measure is closely related to Peterson's (1976) measure of local diversity.

The unbiased estimation theory for $\bar{s}(m)$ is a simple extension of the multinomial case. The set of sample statistics $\mathbf{N}_1 \cdots \mathbf{N}_n$ form a set of complete sufficient statistics for a mixture of multinomial distributions. The MVU estimator for the population mean, $\bar{s}(m)$, is then just the sample mean conditioned on the complete sufficient statistic;

$$\hat{\bar{s}}(m) = \frac{1}{n}\sum_{i=1}^{n} E[S|N_i] = \frac{1}{n}\sum_{i=1}^{n} \hat{s}_i(m). \tag{7}$$

An unbiased estimator for the variance of a single sample estimate $\hat{s}_i(m)$ is

$$\hat{\sigma}_s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\hat{s}_i(m) - \hat{\bar{s}}(m))^2. \tag{8}$$

The estimated variance, $\hat{\sigma}_s^2$, can be partitioned into two components, the estimated variance due to sampling from each multinomial subpopulation,

$$\hat{E}[\mathrm{Var}\,[S|N_i]] = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{V}}\mathrm{ar}\,[\hat{s}_i(m)] \ , \tag{9}$$

and the estimate of the variability in diversity between subpopulations,

$$\hat{\mathrm{V}}\mathrm{ar}\,[s_i(m)] = \hat{\sigma}_s^2 - \hat{E}[\hat{\mathrm{V}}\mathrm{ar}\,[\hat{s}_i(m)]]. \tag{10}$$

Using equation (10), one can evaluate the relative importance of multinomial sampling error and diversity fluctuations within the population. For example, in a typical sampling program a large number of samples are taken but each sample contains only a small number of individuals. In this situation estimates of diversity for a single sample are not accurate, since the sample size is small. The estimation procedure (10) provides a means of assessing the spatial variability in diversity even though the information about diversity in any one sample is limited.

4. Samples from Replicate Communities

We have applied the expected species diversity measure to the diatom data given by Patrick (1968) in her study of replicate diatom communities. This example clearly shows the dependence of the variance of $\hat{\bar{s}}(m)$ on both the size of the original sample and on the individual index, m.

It is clear that samples drawn from natural communities are not random samples from a multinomial population. This is primarily due to the patchy nature of natural communities. The sampling variance derived from the random sampling model may be an underestimate of the true natural variability. As we pointed out in the last section, sampling variance is composed of two parts: the variance due to random sampling and the variance due to variability in the natural communities. When investigating natural communities, the variance

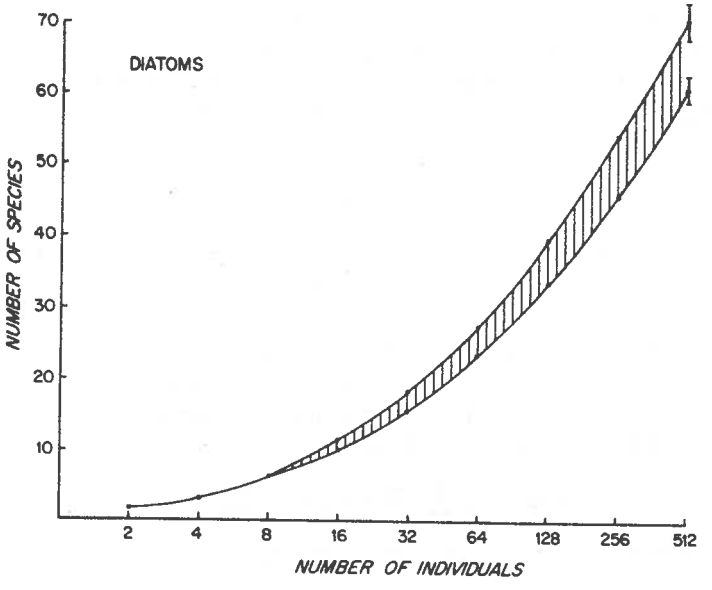

Figure 1
The expected species diversity curves for the eight replicate communities in Patrick's (1968) study. All eight communities fall within the shaded area of the curve. Sampling error for high and low samples is indicated by bars, two standard deviations on each side of the estimates for 512 individuals.

estimates and the confidence intervals derived from them should be thought of as a reasonable lower bounds on the true variability.

Since the sample sizes in Patrick's replicate diatom communities ranged from 3,169 to 6,526, the expected species diversity can be estimated for large values of the individual index, m. Figure 1 gives a curve plotting the expected number of species versus the number of individuals. All eight replicate communities fall within the shaded area of the curve. Because the sample sizes are large the sampling variance of the expected species estimate is small. This yields statistically significant differences between replicate communities even though the numerical differences are rather small.

Figure 2 gives the approximate 95 per cent confidence intervals for the individual indices $m = 2$ and $m = 128$, for all eight communities. The total variation in the diversity estimates for the eight replicate communities can be partitioned into sampling variation and true variability in the diversity index between replicates. Applying (13) we estimate that for the individual index, $m = 2$, 9.6 percent of the total variability is due to sampling error and the rest is due to variations between the replicate communities. For $m = 128$ the variation due to sampling error is only 4.1 percent.

5. Conclusion

The unbiased estimator property of the family of expected species diversities gives this family a number of immediate and practical advantages over other diversity indices in common use. For small field samples the diversity will not, on average, be underestimated; and the variance of the diversity estimate is easily obtained. Also, the variation in the diversity estimates can be simply partitioned into variation due to random sampling error and variation due to spatial heterogeneity in the population.

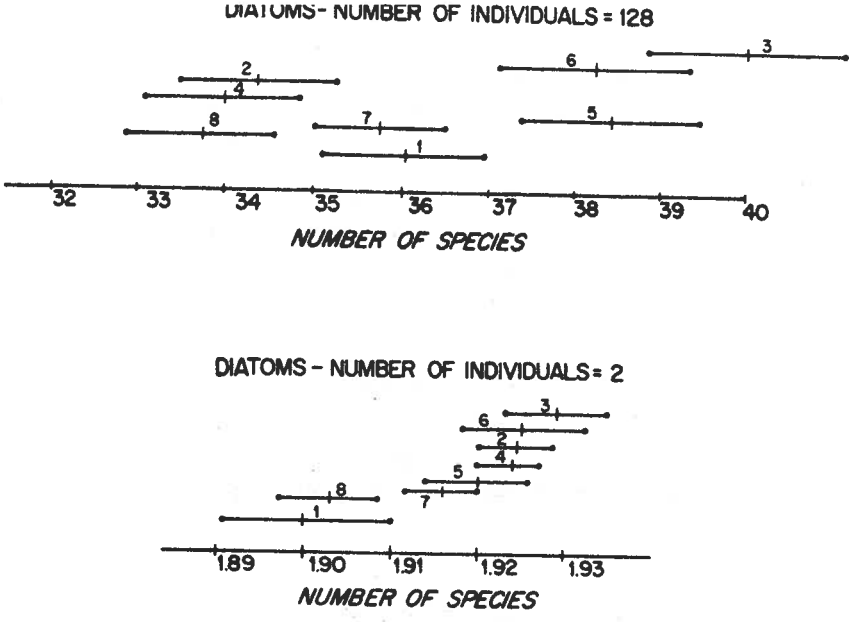

Figure 2
The expected species estimates for 2 and 128 individuals for Patrick's eight replicate diatom communities. The bars are two standard deviations on each side of the estimated value.

Perhaps more important than the relatively good statistical properties outlined is that the expected species index can be used to understand the contribution of rare and abundant species to diversity in natural communities. The expected species index can be estimated for a spectrum of individual indices m. For small m the abundant species dominate the diversity index, while for large m the rare species will also contribute to the diversity index. For gross comparisons between communities diversity could be represented by a curve such as those shown in Figure 1 (Sanders 1968).

Most other diversity measures can be classified as heavily dependent on rare species or abundant species (Peet 1974). The statistical properties of Simpson's measure of diversity are well understood but it is heavily dependent on the abundant species in the population. The expected species diversity can be thought of as the generalization of Simpson's measure to more than two individuals. Shannon's information index, although less dependent on the dominant species than Simpson's index, still gives rare species little weight. Hill (1973) has used the generalized entropy function as a measure of species diversity. By evaluating different points of the generalized entropy function one can emphasize both rare and abundant species. However, it is unlikely that any of the good sampling properties outlined in this paper hold for generalized entropy. May (1975) compares some parametric models for species diversity and notes some problems in evaluating these models when the sample size is small.

Peet (1974), Sanders (1968), Hurlbert (1971) and others have pointed out that the expected species curves for two samples may intersect. Thus one sample would have a higher diversity for, say, $m = 2$ while the other sample would have a higher diversity for $m = 128$. This situation is clearly illustrated in diversity estimates for Patrick's diatom data, Figure 2. Some have used this property of expected species diversity to claim that it could not be used as a diversity index since it gives qualitatively different results for different values of the

individual index m. However, one should remember that expected species diversity is not a single diversity measure but a family of diversity measures based on a single unifying concept. As we pointed out, higher values of the individual index emphasize the rarer species: thus differences in the expected species diversity reflect real differences in the distribution of rare and dominant species. In Patrick's study diversities are so similar that inconsistencies probably reflect real but ecologically insignificant differences in the replicate communities. The overall diversity of the diatom community is appropriately represented by Figure 1.

The value of any ecological index depends not only on its theoretical meaning but also on its statistical properties. Hurlbert's expected species diversity is readily visualized and meets the theoretical and statistical requirements for a diversity measure in a natural way.

Acknowledgments

This work has benefited from a number of helpful discussions with Howard Sanders. Research was supported by NSF Grant GA-36554. We wish to thank George Grice, Loren Haury, Richard Hoffmann, Melvin Rosenfeld and Peter Wiebe for their critical review of the manuscript.

Proprietes Statistiques d'une Famille de Mesures de Diversite

Résumé

On définit une famille de mesures de diversité proposée par Hurlbert comme l'espérance du nombre d'espèces dans un échantillon aléatoire de m individus tirée d'une population. Pour m = 2 cette mesure est équivalente à l'index de diversité de SIMPSON. Tandis que m augmente la mesure devient de plus en plus sensible à la présence d'espèces rares. Dans cet article nous utilisons la théorie de l'estimation non biaisée pour obtenir un estimateur non biaisé de variance minimale pour cette famille de mesures de diversité. Un estimateur non biaisé dans la variance est également obtenu. Ces résultats sont ensuite utilisés pour définir la variation dant la diversité de l'échantillon entre une erreur d'échantillonnage aléatoire et une variation locale de la diversité.

References

Blyth, C. R. (1959). Note on estimating information. *Annals of Mathematical Statistics 30,* 71–79.

Bowman, K. O., Hutcheson, K., Odum E. P, and Shenton, L. R. (1971). Comment on the distribution of indices of diversity. In *Statistical Ecology, Vol. 3.* G. P. Patil, E. C. Pielou and W. E. Waters (eds.), Pennsylvania State University Press, 315–359.

Eberhardt, L. I. (1971). Discussion following: Comments on the distribution of indices of diversity. In *Statistical Ecology,* Vol. 3. G. P. Patil, E. C. Pielou and W. E. Waters (eds.), Pennsylvania State University Press, 359–362.

Fager, E. W. (1972). Diversity: a sampling study. *American Naturalist 106,* 293–310.

Fraser, D. A. S. (1958). *Statistics, an Introduction.* Wiley, New York.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40,* 237–264.

Hessler, R. R. and Jumars, P. A. (1974). Abyssal community analysis from replicate box cores in the central North Pacific. *Deep-Sea Research 21,* 185–209.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology 54,* 427–432.

Hurlbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology 52,* 577–586.

Jumars, P. A. (1975a). Methods for measurement of community structure in deep-sea macrobenthos. *Marine Biology 30,* 245–252.

Jumars, P. A. (1975b). Environmental grain and polychaete species' diversity in a bathyal benthic community. *Marine Biology 30*, 253–266.
May, R. M. (1975). Patterns of species abundance and diversity. In *Ecology and Evolution of Communities*. M. L. Cody and S. M. Diamond (eds.). Belknap Press, Cambridge, Massachusetts. 81–120.
Patrick, R. (1968). The structure of diatom communities in similar ecological conditions. *American Naturalist 102*, 173–183.
Peet, R. K. (1974). The measurement of species diversity. *Annual Review of Ecology and Systematics 5*, 285–307.
Peterson, C. H. (1976). Measurement of community pattern by indices of local segregation and species diversity. *Journal of Ecology 46*, 157–169.
Pielou, E. C. (1969). *An Introduction to Mathematical Ecology*. Wiley-Interscience, New York.
Rao, C. Radhakrishna. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.
Raup, D. M. (1975). Taxonomic diversity estimation using rarefaction. *Paleobiology 1*, 333–342.
Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist 102*, 243–282.
Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon 21*, 213–251.
Williams, C. B. (1964). *Patterns in the Balance of Nature*. Academic Press, New York.

Woods Hole Oceanographic Institution Contribution 3558.
Received May 1976, *Revised August* 1976

Appendix

In this appendix we give detailed steps for computing the sampling variance for the expected species diversity measure. The results given here follow directly from the simple properties of the hypergeometric distribution; however, the calculations become rather complex and time consuming even for large high-speed computers.

The problem is to find an efficient computational form for $E[S' \cdot S''|\mathbf{N}]$ in (6). Let L_i' and L_i'' respectively, denote the number of individuals of species i in the first and second non-overlapping samples of size m; then

$$E[S' \cdot S''|\mathbf{N}] = 2 \sum_i \sum_{j>i} Pr(L_i' \geq 1 \quad \text{and} \quad L_i'' \geq 1|\mathbf{N}) + \sum_i Pr(L_i' \geq 1 \text{ and } L_i'' \geq 1|\mathbf{N}). \tag{A1}$$

The probabilities in the second summation are easy to compute and are written

$$\begin{aligned} f_m(N_i;N) &= Pr(L_i' \geq 1,\ L_i'' \geq 1|\mathbf{N}) \\ &= 2Pr(L_i \geq 1|\mathbf{N}) - Pr(L_i' + L_i'' \geq 1|\mathbf{N}) \\ &= 1 - 2b(0,\ m,\ N_i,\ N) + b(0,\ 2m,\ N_i,\ N). \end{aligned} \tag{A2}$$

where $b(k, m, l, n)$ are the hypergeometric probabilities,

$$b(k, m, l, n) = \frac{\binom{l}{k}\binom{n-l}{m-k}}{\binom{n}{m}} \tag{A3}$$

The probabilities in the first summations in (A4) are not so tractable. The best procedure we can find is the following:

$$\begin{aligned} f_m(N_i,\ N_j;\ N) &= \Pr\,(L_i' \geq 1,\ L_j'' \geq 1|\mathbf{N}) \\ &= 1\text{-}b\ (0,\ m,\ N_i,\ N) - b(0,\ m,\ N_i,\ N) \\ &\quad + \Pr\,(L_i' = 0,\ L_j'' = 0|_{\mathbf{N}}). \end{aligned} \tag{A4}$$

The event $L_i' = 0, L_j'' = 0$ is the union of mutual exclusive events $L_i' = 0, L_j' = k$ and $L_j'' = 0$; $k = 0, 1, 2 \cdots$, thus

$$Pr(L_i' = 0,\ L_j'' = 0|\mathbf{N}) = \sum_{k=\max(0,\ 2m+N_j-N)}^{k\ =\ \min\,(N_j,\ m)} Pr(L_i' = 0,\ L_j' = k,\ L_j'' = 0|\mathbf{N}). \tag{A5}$$

The probabilities on the right hand side of (A5) can be expressed in terms of hypergeometric probabilities

$$\Pr(L_i' = 0, L_j' = k, L_j'' = 0|\mathbf{N}) = \Pr(L_i' = 0|\mathbf{N}) \cdot \Pr(L_j' = k|\mathbf{N}, L_i' = 0) \cdot \Pr(L_j'' = 0|\mathbf{N}, L_i' = 0, L_j' = k)$$
$$= b(0, m, N_i, N) \cdot b(k, m, N_j, N - N_i) \cdot b(0, m, N_j - k, N - m). \qquad \text{(A6)}$$

This formula computationally long and can be shortened by using the iterative formula

$$Pr(L_i' = 0, L_j' = k + 1, L_j'' = 0|\mathbf{N}) = Pr(L_i' = 0, L_j' = k, L_j'' = 0|\mathbf{N}) \cdot \frac{(N_j - k) \quad (m - k) \quad (N - m - N_j + k + 1)}{(k + 1)(N - N_i - N_j - m + k + 1) \quad (N - 2m - N_j + k + 1)}.$$

We can further shorten this procedure by noting that for every pair of species with l and l' individuals, respectively, the computations are identical. Let k_l denote the number of species with exactly l individuals in the sample. The final form of equation (A1) is then

$$E[S' \cdot S''|\mathbf{N}] = \sum_l \sum_{l'>l} 2 \cdot k_l \cdot k_{l'} \cdot f_m(l, l', N) + \sum_l k_l \cdot (k_l - 1) f_m(l, l, N) + \sum_l k_l f_m(l, N)$$

All computations should be carried out in double precision. The hypergeometric probabilities, $b(k, m, l, h)$ are computed in double precision using the International Mathematical and Statistical Libraries, Inc. subroutine MDHYP. A subprogram that computes $\hat{s}(m)$, $\widehat{\text{Var}}[S]$ and $\widehat{\text{Var}}(\hat{s}(m))$ using these results is available from the first author.

Reprinted from
BIOMETRICS COPYRIGHT © 1977
THE BIOMETRIC SOCIETY, Vol. 33, No. 2, June 1977