*Ecology* (1975) **56**: pp. 1459–1461

# EXPLICIT CALCULATION OF THE RAREFACTION DIVERSITY MEASUREMENT AND THE DETERMINATION OF SUFFICIENT SAMPLE SIZE[1]

KENNETH L. HECK, JR.
*Department of Biological Science, Florida State University, Tallahassee, Florida 32306 USA*

GERALD VAN BELLE
*School of Public Health and Community Medicine, Department of Biostatistics SC-32,
University of Washington, Seattle, Washington 98195 USA*

DANIEL SIMBERLOFF
*Department of Biological Science, Florida State University,
Tallahassee, Florida 32306 USA*

*Abstract.* An explicit means of calculating the expected number of species [$E(S_n)$] and the variance of ($S_n$) in a random sample of $n$ individuals from a collection containing $N$ individuals and $S$ species is presented. An example illustrates a new use of $E(S_n)$: determination of the sample size required for any desired degree of accuracy in collecting species known to occur in a particular area.

*Key words*: *Expected number of species*; *rarefaction*; *sample size.*

## INTRODUCTION

Hurlbert (1971), in a review of the rapidly proliferating literature on species diversity, concluded that most of the commonly used indicators of biotic diversity, including those based on information theory, are inappropriately applied to biotic collections and functionally meaningless as descriptions of biological properties. One of two alternatives which he proposed is $E(S_n)$, the expected number of species in a sample of ($n$) individuals selected at random from a collection containing ($N$) individuals and ($S$) species. In this method the number of species (species richness) is calculated for the collections to be compared after all collections are scaled down to the same number of individuals (presumably that in the smallest collection). This scaling is necessary because large collections would have more species than small ones even if they were drawn from the same community. The situations where we wish to determine whether two samples could conceivably have been drawn from the same community, or whether a small sample can conceivably be considered a random subsample of a large one, may also be dealt with using the same statistic (Simberloff 1970).

In this communication we first restate Hurlbert's explicit means of calculating $E(S_n)$, then present an equation for $\mathrm{var}(S_n)$; finally we show an additional use of $E(S_n)$, to determine the sample size required for any desired degree of accuracy in collecting species known to occur in a particular area.

## EXPLICIT CALCULATION OF $E(S_n)$ AND VAR($S_n$)

Sanders (1968) used the statistic $E(S_n)$ in a major work on marine benthic diversity, calling his scaled down samples "rarefied" samples and the scaling procedure "rarefaction," but did the scaling incorrectly; his method consistently overestimated the number of species which would be expected in a sample of ($n$) individuals drawn randomly from the larger collection (Hurlbert 1971, Fager 1972, Simberloff 1972). The general reasons for the overestimation were given by Hurlbert (1971) and Simberloff (1972), who showed that the true $E(S_n)$ is determined by the hypergeometric distribution; the latter author wrote a simple computer program to simulate the random drawing any number of times and thus to estimate not only the true $E(S_n)$ but also the confidence limits about this parameter.

The program is no longer necessary, since $E(S_n)$ and $\mathrm{var}(S_n)$ can both be calculated directly. If $N_i$ is the number of individuals in species $i$ of the unrarefied sample, then

$$(1) \qquad E(S_n) = S - \binom{N}{n}^{-1} \sum_{i=1}^{S} \binom{N-N_i}{n} \qquad \text{and}$$

$$(2) \quad \mathrm{var}(S_n) = \binom{N}{n}^{-1} \left[ \sum_{i=1}^{S} \binom{N-N_i}{n} \left( 1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right) + 2 \sum_{\substack{j=2 \\ i<j}}^{S} \left( \binom{N-N_i-N_j}{n} - \frac{\binom{N-N_i}{n}\binom{N-N_j}{n}}{\binom{N}{n}} \right) \right].$$

If ($n$) is so much smaller than $N$ that drawing $n$ individuals randomly can be approximated by sampling with replacement, or if there are biological

[1] Manuscript received 26 April 1974; accepted 11 August 1975.

reasons why the successive drawings of individuals for the smaller sample can be construed as not affecting the species-individuals distribution of the original, larger sample, then the multinomial distribution applies instead of the hypergeometric and

$$(3) \qquad E(S_n) = S - \sum_{i=1}^{S} (1 - N_i/N)^n \qquad \text{and}$$

$$(4) \quad \operatorname{var}(S_n) = \sum_{i=1}^{S} (1 - N_i/N)^n [1 - (1 - N_i/N)^n] + 2 \sum_{\substack{j=2 \\ i<j}}^{S} [(1 - N_i/N - N_j/N)^n - (1 - N_i/N)^n (1 - N_j/N)^n].$$

Formulae (3) and (4) can be found in Harris (1959).

It should be apparent that the statistical expectation of the original random draw method of Simberloff (1972) and the direct calculations given above are identical. A computer program which calculates means and variances under both of the above sampling schemes, given the original distribution of $N$ individuals among $S$ species, is available upon request from the authors.

## Use of $E(S_n)$ in the Determination of Sufficient Sample Size

We now consider some of the problems associated with a commonly used method of deciding on a sufficient sample size, and show the potential use of $E(S_n)$ in overcoming these problems.

One of the means employed most often to determine sample size is to plot the number of new species obtained per unit sample as a function of sample number; an asymptote is taken to indicate sufficient sample size. A *caveat* concerning this method should be recognized however: in a very patchy environment the species accumulation curve may become asymptotic before many of the species have been sampled. This problem can be overcome if time and other considerations allow samples to be taken until several successive replications produce no new species. More important, though, is the problem which arises when the sample size required to approach the asymptote on the species accumulation curve is too large to be practical. Then the difficulty occurs because the relationship between sample size and the accumulation of new species is not linear (Preston 1960, 1962). The following example presents an application of the rarefaction diversity measure that eliminates the problem of the nonlinear relationship between sample size and the accumulation of new species; we show a method of determining the sample size required for any desired degree of accuracy in estimating species presence. What is required to set the degree of accuracy of a sampling scheme is a large sample taken in an area of interest, whose total number of species ($S$) and distribution of individuals among species is known. This original large sample must be large enough so that an almost complete enumeration of the species present has been obtained. Then, by using the formulae given above for calculating $E(S_n)$ and $\operatorname{var}(S_n)$ one can select a sample size ($n$) which allows, on the average, a desired proportion of ($S$) to be taken.

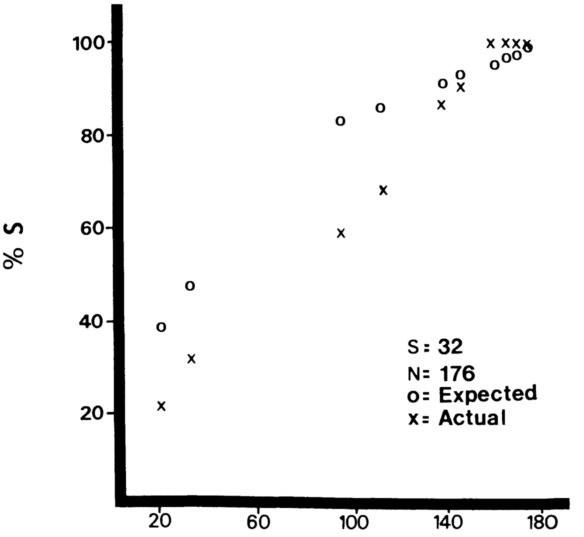

FIG. 1. The expected [$E(S_n)$] and actual species accumulation curves for 1 mo at one sampling site. Values are expressed as percentages of the total number of species ($S$) obtained in the samples. Sample size represents the cumulative number of individuals obtained after each of 10 randomly taken trawl samples of identical duration. The individual sample sizes were: 21, 11, 68, 16, 25, 7, 16, 4, 4, and 5. These data indicate a somewhat clumped distribution of species but the difference between actual and expected points is not statistically significant, as determined by the binomial test (cf. Hessler and Jumars 1974). This plot is representative of all monthly plots from the four sampling sites.

This procedure will be most effective if the original large sample is broadly inclusive both temporally and spatially, so that the effects of seasonality and beta diversity on species number and the relative abundance of species can be minimized. If there is little temporal variability in the relative abundances of the taxa of interest then one could use the entire large sample to determine an appropriate sample size. Since this will probably only rarely be true, we suggest that data be divided into seasonal, monthly, or other temporal subunits, according to rates of fluctuation observed in the relative abundances of the taxa being studied, and appropriate estimates of sample size be determined for each of the temporal subunits. In the following example we calculated monthly estimates of sufficient sample size because relative abundances of the taxa of interest has been found to fluctuate from month to month (Hooks 1973).

## Example

Data gathered by Hooks (1973) over a 12-mo period were available for analysis before a second study was begun in the same area. The data consisted of species counts of benthic macroinvertebrates taken by repeated trawling at four intensively sampled sites off Florida's north-central coast which had been sampled monthly until the species accumulation curve became asymptotic. That an asymptote was reached at all sites suggests that most or all of the species were included. These data also showed that species frequency distributions became nearly constant well before the species accumulation curve. Since additional sampling sites were to be investigated in the second study it was necessary, in terms of cost, to reduce sampling intensity. Analysis of the original raw data using the rarefaction formulae (1) and (2) showed that in nearly every month at each of the four stations an $E(S_n) \geqslant 90\%$ of $(S)$ could be obtained with only 70% of the original sampling effort (Fig. 1). Thus, the total number of sampling sites was increased without an unacceptable (to us) loss of information at the existing sites.

This method of determining sample size has potential relevance to any long-term biotic sampling program. After a thorough and exhaustive initial survey, sample size may often be decreased without a serious loss of information. In some situations one might be satisfied to collect, on the average, only 50%–75% of the total number of species known to occur in a given area as long as the most common species were obtained. In other situations a study might not be continued, or a follow-up study begun, if the method showed that the greatest sustained sampling intensity affordable would only sample an unacceptably small proportion of the total species pool.

In conclusion, it should be emphasized that the procedure outlined above is in no way specific for one type of sampling apparatus, but is broadly applicable to samples taken with any type of sampling gear, and for any taxocene, provided that the assumptions upon which the sampling schemes themselves are based have been satisfied.

## Literature Cited

Fager, E. W. 1972. Diversity: A sampling study. Am. Nat. **106**:293–310.

Harris, B. 1959. Determining bounds on integrals with application to cataloging problems. Ann. Math. Stat. **30**:521–548.

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

Hooks, T. A. 1973. M.S. Thesis. Florida State Univ., Tallahassee.

Hurlbert, S. H. 1971. The non-concept of species diversity: A critique and alternative parameters. Ecology **52**:577–586.

Preston, F. W. 1960. Time and space and the variation of species. Ecology **41**:611–627.

———. 1962. The canonical distribution of commonness and rarity. Pt. I. Ecology **43**:185–215.

Sanders, H. L. 1968. Marine benthic diversity: A comparative study. Am. Nat. **102**:243–282.

Simberloff, D. S. 1970. Taxonomic diversity of island biotas. Evolution **24**:23–47.

———. 1972. Properties of the rarefaction diversity measurement. Am. Nat. **106**:414–418.