Theoretical Population Biology **54**, 294–302 (1998)
Article No. TP971370

TPB

A General Theory of the Sampling Process with Applications to the ''Veil Line''

A. K. Dewdney

Department of Computer Science & Department of Zoology,
The University of Western Ontario, London, Ontario, Canada N6A 5B7

Received July 22, 1997

When a community of species is sampled, nonappearing species are not those with abundances that fall shy of some arbitrary mark, the ''veil line'' proposed by E. F. Preston in 1948 (*Ecology* 29, 254–283). Instead, they follow a hypergeometric distribution, which has no resemblance to the veil line. There is therefore no justification for the truncation of distributions proposed to describe the abundances of species in natural communities.

The mistake of the veil line points to the need for a general theory of sampling. If a community has a distribution *g* of species abundances and if samples taken of the community tend to follow distribution *f*, what is the relationship of *f* to *g*? The seeds of such a theory are available in the work of E. C. Pielou. Using the Poisson distribution as a close approximation to the hypergeometric, one may immediately write and (in most cases) solve the transformation from *g* to *f*. The transformation appears to preserve distribution formulas to within constants and parameters, providing yet another reason to rule out the use of truncation. Well beyond this application, the theory provides a foundation for rethinking the sampling process and its implications for ecology. © 1998 Academic Press

INTRODUCTION

In a community, each species at any moment will have a particular abundance. When the species are sorted into abundance categories from one to the maximum abundance, a species abundance distribution emerges as a discrete histogram. By the same token, if the community is sampled and the species of the sample are sorted in the same manner, another species abundance distribution emerges as another histogram. In general, we have no direct information about the kind of abundance histograms that prevail in actual communities. We have only samples of those communities.

A theory that describes the patterns of species abundances in communities might take one of two routes: It might propose a particular theoretical distribution as one that underlies all abundance patterns, then attempt to establish it through the analysis of existing biosurveys. This happened in the 1940s, for example, when the American ecologist E. F. Preston (1948) proposed the lognormal distribution for just such a role. It happened again when the British entomologist C. B. Williams, who was never convinced of the validity of the lognormal distribution, proposed the log-series distribution (1964) in its place. Over the past 50 years, we find not only the lognormal and the log-series distributions, but the negative binomial (Pielou, 1975), the Zipf–Mandelbrot (Zipf, 1965; Mandelbrot, 1977), and the logistic distributions (Dewdney, 1997) among others.

Another theory of species abundances might begin more generally by examining the mathematical relationship between the abundance distribution *g* in a community and the abundance distribution *f* of its samples. Such a theory would make no reference to a particular proposal, but would seek instead to establish a general relationship between *f* and *g*. But such a theory might rule out a proposed distribution in an *a priori* sense.

The plethora of species abundance distributions (Magurran, 1988) proposed over the past 50 years

0040-5809/98 $25.00

Copyright © 1998 by Academic Press
All rights of reproduction in any form reserved.

speaks not to the development of the field of mathematical ecology, but rather to its confusion. Intellectual gymnastics are required (May, 1975) to give each of them a role in the various kinds of communities. In the present circumstances, if it should be the case that only one universally valid species abundance distribution actually exists, we would never learn of it. It is proposed here that a truly general theory of community sampling is within reach and should be developed. The outlines of such a theory emerge from the work of Pielou (1969).

Both the need for such a theory and for its application are demonstrated by the notion of the "veil line" (Preston, 1948). In that year, Preston proposed that the distribution of species abundances follows the lognormal distribution, essentially a log-transformed normal distribution. This (continuous) function begins with value 0 at the origin, then rises in an almost normal-curve fashion to a maximum before falling again toward the high-abundance end of the axis. One can visualize the lognormal readily enough by imagining a normal distribution that is stretched outward by amounts that systematically double. Figure 1 shows a typical lognormal pdf.

In order to compare the lognormal distribution with a real sample we may avoid the awkwardness of the distribution formula by transforming the sample instead of the normal distribution. The inverse transformation is exponential, compressing the sample inward by adding up all the species that fall within intervals that are themselves doubling in size toward the high end of the abundance axis. Following this transformation, the sample can be compared directly with the appropriate (truncated) normal distribution.

Preston proposed the lognormal as the prevailing distribution in biological communities, but was forced to establish the hypothesis through an examination of several samples. None of the samples showed a marked resemblance to the (untruncated) lognormal distribution. Nevertheless, it has become standard practice for some biologists to accept the lognormal distribution and its accompanying veil line at face value.

Some of the distributions examined by Preston had the shape of the "J-curve," a term that does not seem to appear in the literature but which some biologists use informally. It describes a typical species abundance (sample) histogram that begins at the low abundance end with the greatest number of species, then ramps rapidly down like the tail of the letter J (reversed) toward the high abundance end of the histogram. Most field biologists have witnessed this general phenomenon of samples (Pielou, 1969): The most abundant species are themselves few in number. The least abundant species are the most numerous.

In the face of this discrepancy, the lognormal would have died a natural death were it not for the "veil line," a straight, vertical line which amounts to the rudiments of a general theory of Preston's. In his opinion (Preston, 1948, 1962), the relationship between a community distribution g and a sample distribution f is very simple. If g is simply truncated it becomes f after an appropriate change of parameter value(s). In other words, all species to the left of a certain veil line in the community distribution

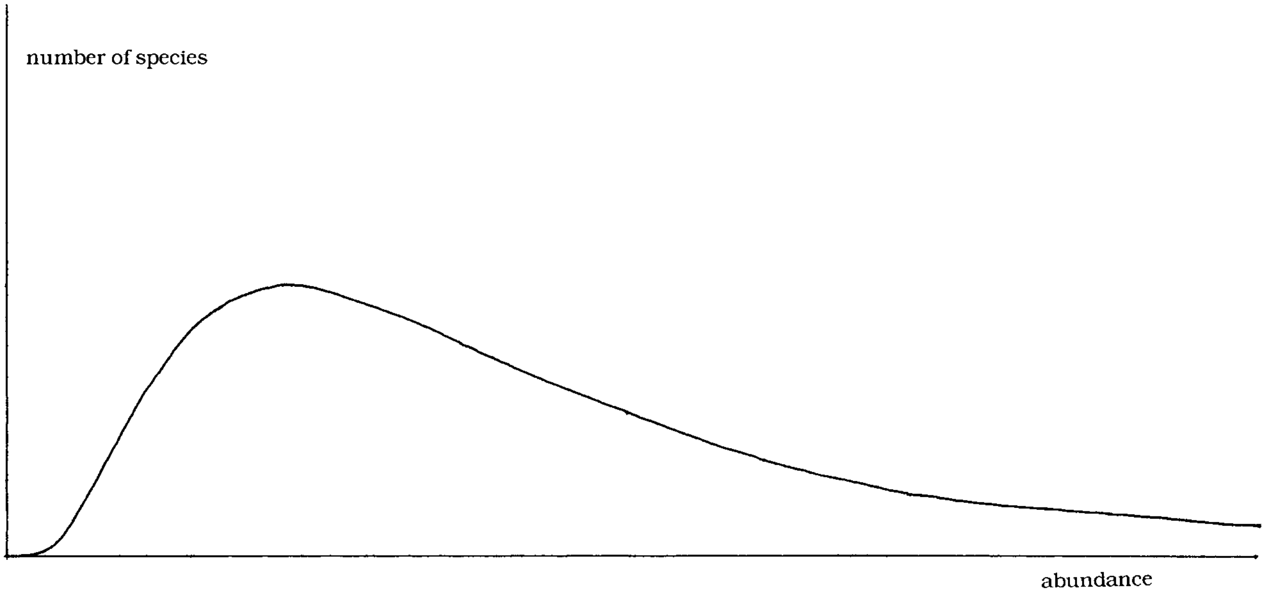

FIG. 1. The lognormal distribution.

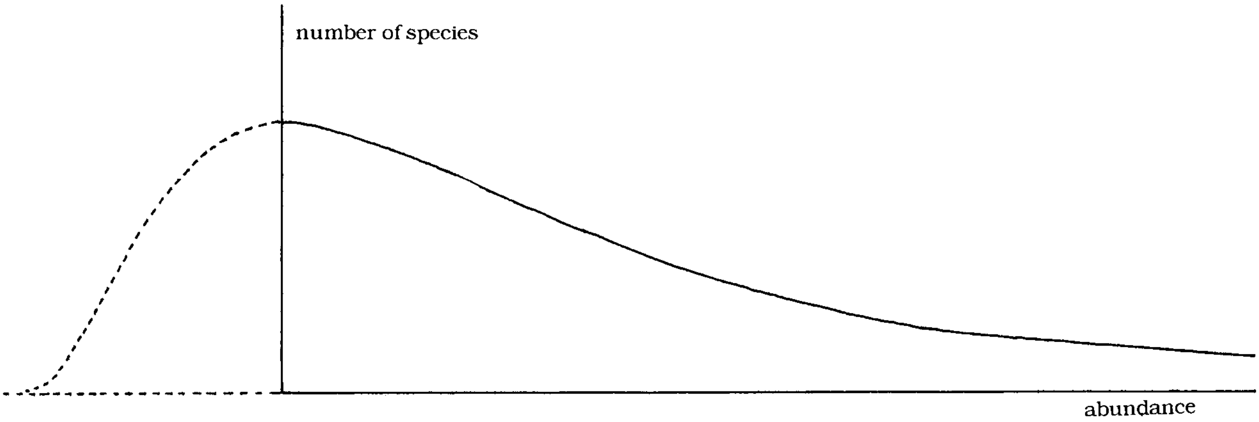

FIG. 2. The truncated lognormal distribution.

simply fail to show up in the sample. Figure 2 shows the lognormal of Fig. 1 after a truncation operation that is typical of applications (Magurran, 1988).

It will be immediately observed that truncation removes the embarrassing dip at the low end of the distribution, resulting in a curve that more closely resembles the J-curve.

If the truncation operation is not permissible, what should we make of the lognormal distribution as a valid model of species-abundance distributions? As we are about to see, the *actual* truncation operation makes the lognormal distribution even less like the J-curve than the original.

VANISHING SPECIES

When biologists sample a given community, they generally do not gather or observe all individuals of the community, but rather a subset of them. With obvious exceptions (such as the trees in a smallish woodlot) a complete sample is usually impossible to take. In what follows we will assume that all samples are unbiased random samples.

Let a community consist of N individuals distributed among m species with abundances $n_1, n_2, n_3, ..., n_m$. If an individual is drawn at random from the community, the probability that it belongs to the ith species will be ni/N. If a random unbiased sample of size n is taken of the community, the number of individuals from species i expected to show up in the sample will therefore be nn_i/N. However, this many individuals will not always show up. The actual distribution is governed by the hypergeometric function (Feller, 1957). The probability that the ith species will contribute k individuals to a sample is given by the following expression:

$$p(k) = \frac{\binom{n_i}{k}\binom{N-n_i}{n-k}}{\binom{N}{n}}, \qquad 0 \leqslant k \leqslant n_i. \tag{1}$$

This expression is readily proven to be correct by an appeal to the standard urn model of sampling. Moreover, this expression is not approximate, but exact. It follows that if there are $g(j)$ species of abundance j in the community and repeated samples of size n are taken (with replacement), the numbers of individuals from such species to show up in the sample will have the following distribution:

$$f(k, j) = t(k, j) \cdot g(j),$$

where

$$t(k, j) = \frac{\binom{j}{k}\binom{N-j}{n-k}}{\binom{N}{n}}, \qquad 0 \leqslant k \leqslant j. \tag{2}$$

Thus, if 100 samples of size n are taken from the community, for example, and we knew all the numbers involved, the only difference between the distribution of values of $f(k, j)$ that might be observed and those predicted by expression (2) will be accounted for by the usual statistical fluctuations that accompany all samples.

The function t will be called the *sample transformation function.* A useful and important property of expression (2) is that the community distribution g and the sample transformation function t are uncoupled. This enables us to apply the transformation without alteration to any community distribution g that we wish. Figure 3 displays

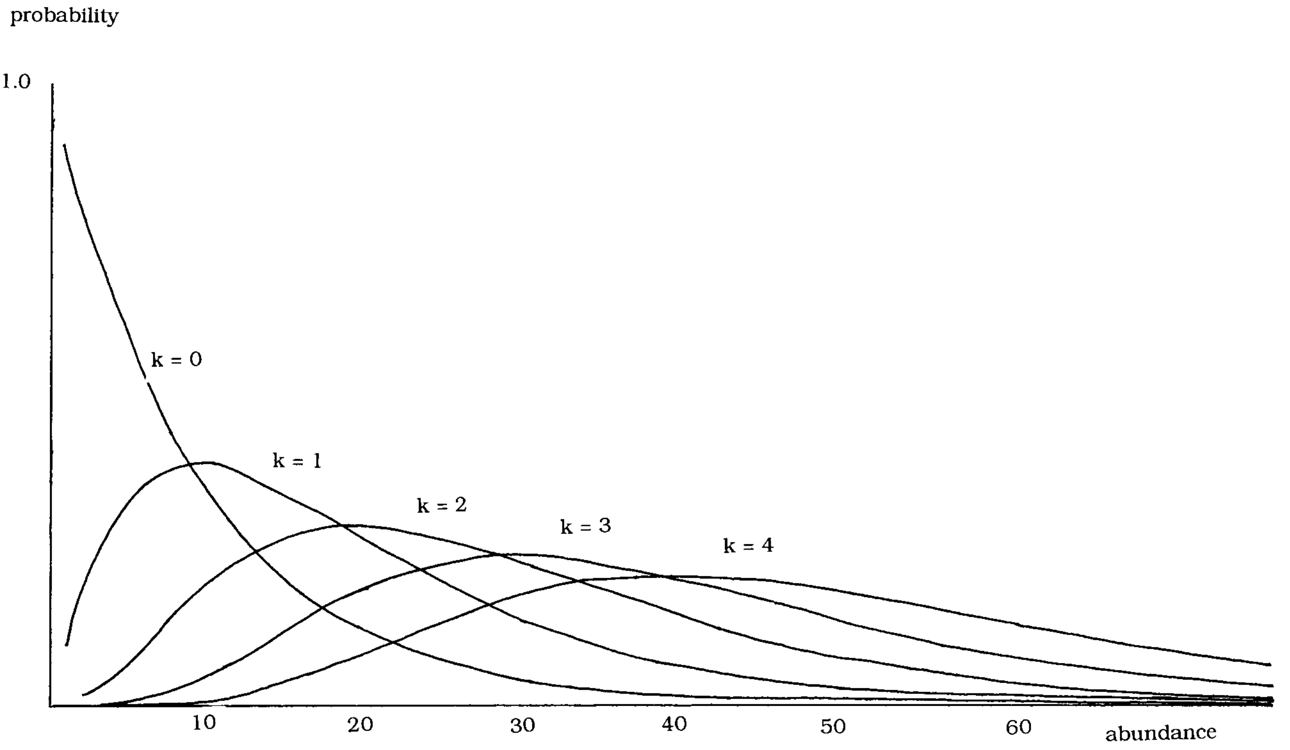

FIG. 3. Sample transformation functions.

an example of the sample transformation functions when $N = 1000$ and $n = 100$. The first few have been plotted on an axis that represents abundance in the community. It may be seen that, with the exception of the 0 class, contributions to each abundance class rise from zero, reach a maximum, and then gently decline.

To gain some insight into how the sample transformation functions operate in an actual sample, we might take g to be the uniform distribution, as shown in Fig. 4. In this example the number of species in each abundance class in the community will contribute differentially to the abundance classes in the sample according to the

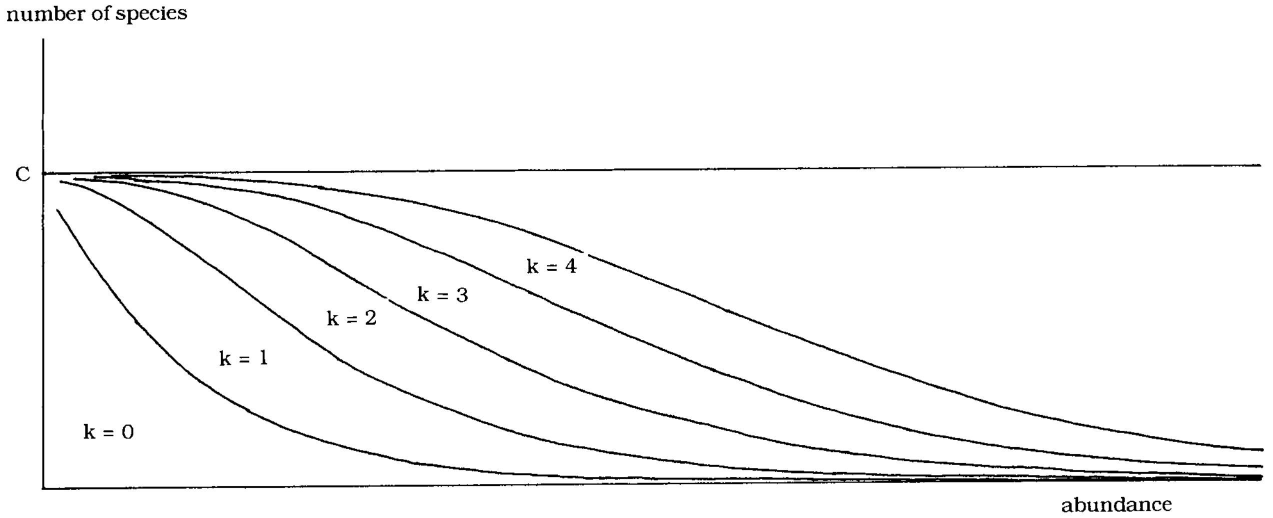

FIG. 4. How the transformation works.

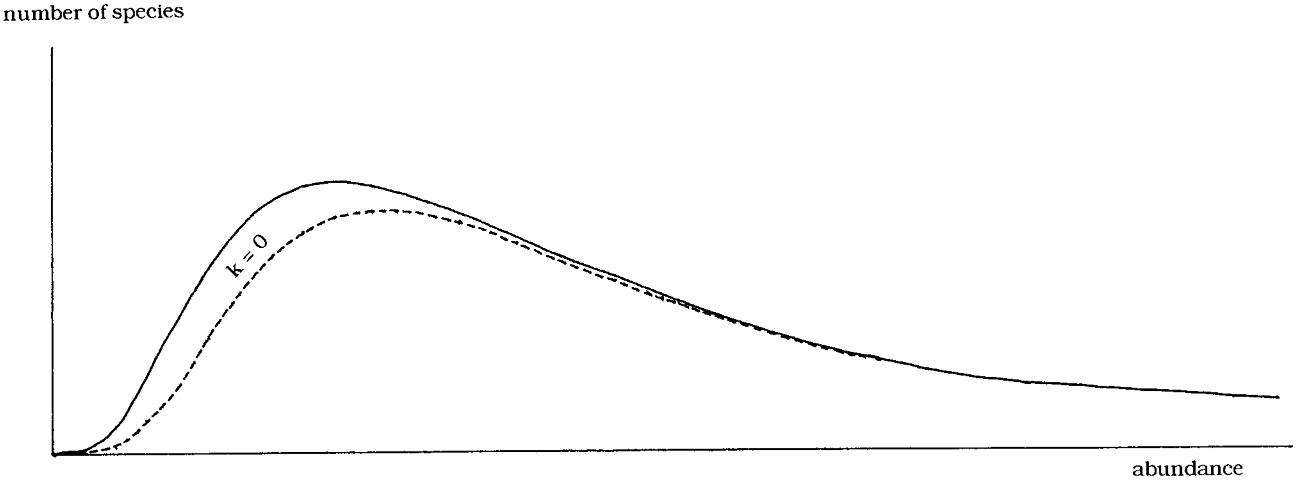

FIG. 5. The veil curve for the lognormal distribution.

partition of the number by the sigmoid-shaped areas that represent the sample transformation functions. The number of species to show up k times in the sample is given by the area of the sigmoid bearing that numerical label.

The zeroth-order transformation function thus represents the species destined to disappear from the sample (to within the usual statistical fluctuations). The boundary of this curve represents the "veil line" for the distribution g. In the standard statistical sense all of the species within this region will not show up in the sample. It is clearly not a vertical straight line, but rather a sloping, curved one.

With the apparatus developed to this point, we may now demonstrate what happens when abundances in the community have a lognormal distribution. Figure 5 shows the lognormal curve for which the zeroth-order class has been computed for the case at hand ($N = 1000$, $n = 100$).

No matter how large N and n are, or how small n is in relation to N, the same shape emerges on every occasion. Those who would use the lognormal distribution may not truncate it with a vertical line, but with the appropriate version of the curvilinear one that appears in the figure. When the zero-order sigmoid is excised from the distribution, what remains looks not more, but rather less, like the J-curve that prevails in samples.

A GENERAL THEORY OF SAMPLING

A general theory of sampling, as proposed in the Introduction, ought to enable us to make some general statement about the relationship between the distribution g that prevails in a community and the distribution that arises in samples of it. The compound formula (2) provides just such a link between g, the community distribution, and f, the sample distribution. The computation of $t(k, j)$, however, presents some problems. Partly due to its discrete nature and partly due to the awkwardness created by the factorial elements in the hypergeometric distribution, it is somewhat unwieldy to work with directly.

Fortunately, Pielou (1969) provides the ingredients for a continuous version of the theory that is much easier to work with. The hypergeometric function is closely approximated by the Poisson distribution (Feller, 1957). Specifically, in the expression (2), let n become arbitrarily large while $j \cdot n/N = \lambda$ remains constant. In the limit, this expression converges to the following function:

$$e^{-\lambda}(\lambda)^k/k!$$

At this point it is appropriate to clean up our notation slightly, substituting for the ratio n/N another constant, r, which we will call the *sampling ratio*. In other words, we may allow ourselves to write the following equality with the understanding that it is only an approximation:

$$p(k) = e^{-jr}(jr)^k/k!$$

Here, j is an arbitrary abundance class in the community distribution. As Fig. 6 demonstrates in the case of our ongoing example, the Poisson distribution is nevertheless a close approximation to the hypergeometric, even for low values of k.

It follows that the discrete transformation of expression (2) may now be replaced by an integral equation

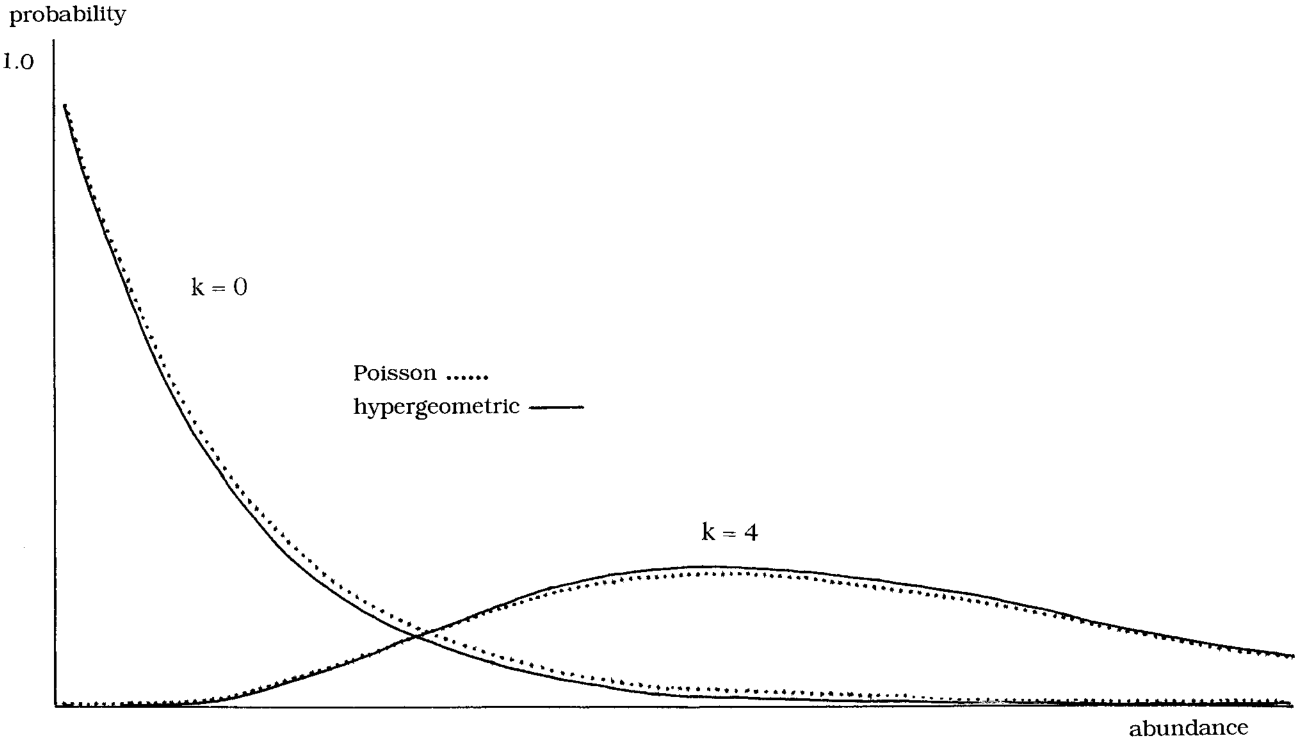

FIG. 6. The Poisson distribution approximates the hypergeometric.

that closely approximates the action of Eq. (1) incorporated into a sum over j-values:

$$f(k) = \int_0^{\infty} (e^{-rx}(rx)^k/k!) \cdot g(x)\, dx. \tag{3}$$

Equation (3) amounts to a transformation of g which we will call the *Pielou transformation*. Note that the discrete variable j has been replaced by the continuous variable x which has the interpretation of an abundance. Like its discrete counterpart, this equation adds up the contributions of the species in each abundance class x to the species that will inhabit the kth abundance category in the sample. This equation may be put to immediate use by calculating in advance the effect of sampling on various distributions g.

The effect is calculated by performing the integration in (3), if possible. In particular cases this may sometimes be accomplished by taking constants outside the integral sign so as to leave an integral of the form

$$\int_0^{\infty} (e^{-rx}(rx)^k/k) \cdot dx.$$

No matter what value k has, this integral always equals $1/r$. This fact is very useful in the solution procedure just described. It may be applied, for example, to the uniform abundance distribution. If we let g have the same number c of species in each abundance category out to some limit, L, elementary manipulations of the Pielou transformation readily yield the following form of f:

$$f(k) = c/r.$$

In other words, f is also a uniform distribution that has been compressed laterally by the factor $1/r$. If the largest abundance in the community were L, then the largest abundance in the sample would be rL, neglecting a small "tail" of even larger abundances.

A number of proposed species abundance distributions are locally inverse linear (or hyperbolic), being proportional to $1/x$. These include the log-series distribution of Williams (1964) and Fisher (1943), the Zipf–Mandelbrot distribution (Zipf, 1965; Mandelbrot, 1977), and the logistic distribution (Dewdney, 1997). Apart from multiplicative or additive factors that force convergence in either the finite or infinite sense, such functions have the form $g(x) = c/x$. When the integration of (3) is performed for this class of functions, a member of the same class invariably results.

$$f(k) = c/k$$

When g is the lognormal function, the resulting transformation turns out to be nonintegrable in the same sense as the normal distribution. However, there are many functions with the same general shape as the lognormal, including the Poisson distribution itself. When $k = 1$, for example, this distribution defines the 1-class in Fig. 3. If we set

$$g(x) = c_1 x \cdot e^{-c_2 x},$$

for example, and apply the Pielou transformation (3) we obtain the same functional form, to within constants.

$$f(k) = c'_1 k \cdot e^{-c'_2 k}$$

These transformations all illustrate, but do not establish, that for a wide class of distribution functions the sampling transformation preserves the mathematical form of the function and therefore its basic shape. Is it possible to say more about this class? Given the nonintegrability of some functions, there are two possible steps to establishing the theory on a firm footing.

First, it may readily be verified that the transformation of a polynomial (in x) of degree n results in another polynomial (in k), also of degree n. It follows that a polynomial approximation (Froberg, 1985) to a distribution function will have its general mathematical form preserved by the Pielou transformation.

Second, since polynomials are preserved by the Pielou transformation, so is any power series expansion of an arbitrary analytic function. To apply this idea successfully, the domain of an analytic function may have to be broken up into finite intervals of convergence.

On these bases, one may conjecture that the Pielou transformation preserves the form of any analytical function. By "form" we mean that the formula of the sample distribution will differ from the formula of the parent distribution only in the value of its constants.

Although the form (and basic shape) of the lognormal distribution is probably also preserved by the Pielou transformation, the situation is complicated slightly in actual calculations. If one plots the values of a typical transform of the lognormal function, the resulting histogram will not show quite the same shape, but will *appear* to be slightly truncated. This phenomenon is due partly to the abrupt rise in the distribution at the low abundance end of its domain and partly to the relatively large area of the sigmoids into which sampling implicitly partitions community species. In other words, the lowest abundance category of the sample may have nearly as many species as the second category because it has accumulated not only the few low abundance species, but many species of higher abundance, as well.

What might be called the "sigmoid effect" applies not just to the lognormal but to all possible g-distributions equally. It is only visible in some distributions, owing to their shape. The inverse linear class of functions, for example, shows no apparent effect of the transformation in the formula that results. It has the same constant c as it had before. The first-order sigmoid of the g-distribution has an area that equals the number of species in the first interval implicit in g.

The sigmoid effect must not be confused with the truncation implicit in Preston's veil line concept. The former phenomenon affects all distributions whatever. The only reason for truncating the lognormal distribution before even starting is that it does not work without truncation.

BIOSURVEYS FROM THE LITERATURE

As part of a concerted effort to analyze natural community abundance distributions, the author has, to date, examined some 50 randomly selected biosurveys from the literature (Dewdney, 1998), with another 50 scheduled to be examined before the project is complete. The surveys span a large range of organisms, from microbiota to fish, insects, birds, and plants. They cover biomes from one pole to the other and they include terrestrial, freshwater, and marine habitats. None of the species-abundance distributions resemble an untruncated lognormal distribution. In fact,very few of them even resemble a truncated one. All could readily be described as J-curves, however.

One must assume that Preston had the misfortune to use atypical examples in formulating his theory of the lognormal distribution and its accompanying veil line.

Preston's original question might nevertheless be asked anew. Why do so many sample distributions resemble a truncated normal distribution when subjected to the transformation by octaves, the exponential transformation described near the end of the introduction to this paper. The result often resembles a normal curve that has been cut more or less in half.

One way that such a distribution can arise from this process is illustrated by the *logistic distribution* of the author (Dewdney, 1997). The pdf of this distribution has a relatively simple form:

$$g(x) = c(1/x - \delta), \qquad \varepsilon \leqslant x \leqslant \varDelta.$$

Here, c is a normalizing constant, ε and $\varDelta$ are the logistic parameters, and $\delta = 1/\varDelta$. If one applies the exponential

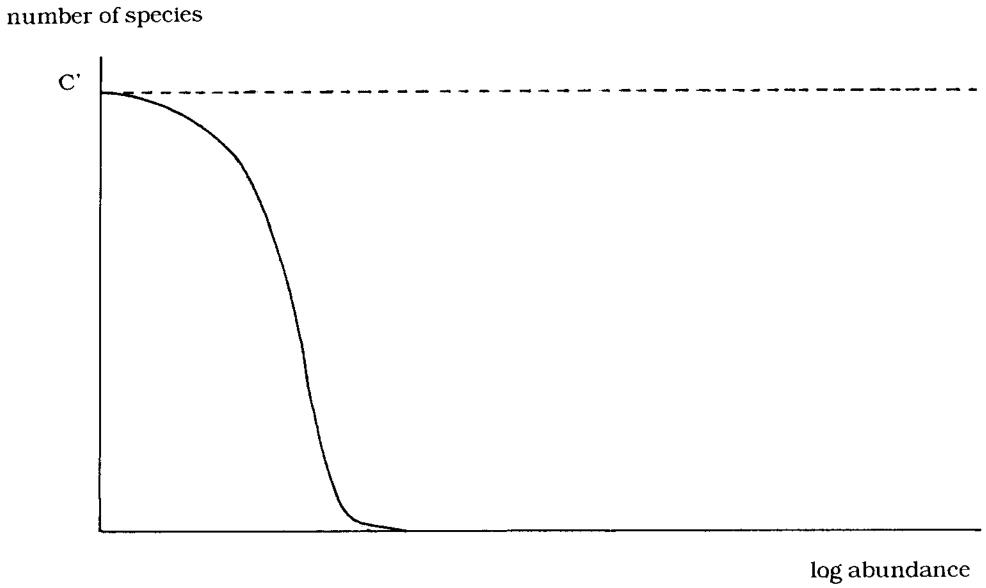

FIG. 7. The exponentially transformed logistic function.

transformation to this function, the resulting distribution resembles a truncated normal distribution. (See Fig. 7 above.) The resemblance arises chiefly from the subtraction of the term δ, continuously doubled, from a constant c', out to the logistic limit Δ.

$$c' - 2^x\delta$$

Given the great tolerance of goodness-of-fit tests, such a shape, as manifested in a sample, would pass for a truncated normal in most cases.

It may be that part of the motivation behind the proposal of truncation was an aversion to the idea that a sample (let alone a community) should have so many species in low abundance. The author was at one time in full sympathy with this view until the realization dawned that a great many species of low abundance did not mean that they were all due for local extirpation, let alone extinction. The illusion vanishes when one considers the likely appearance of an abundance distribution in a reasonably large community of organisms. Instead of the compact shapes this paper displays, a privileged glimpse of an actual distribution is more likely to reveal something like the pattern shown in Fig. 8.

One cannot tell, merely by looking at this pattern, what distribution the abundances are likely to have. Sampled with a ratio of something like 0.001, however, this might easily be the parent for a typical J-curve distribution. Few of the constituent populations appear to be at risk and the ones with lowest abundance might well

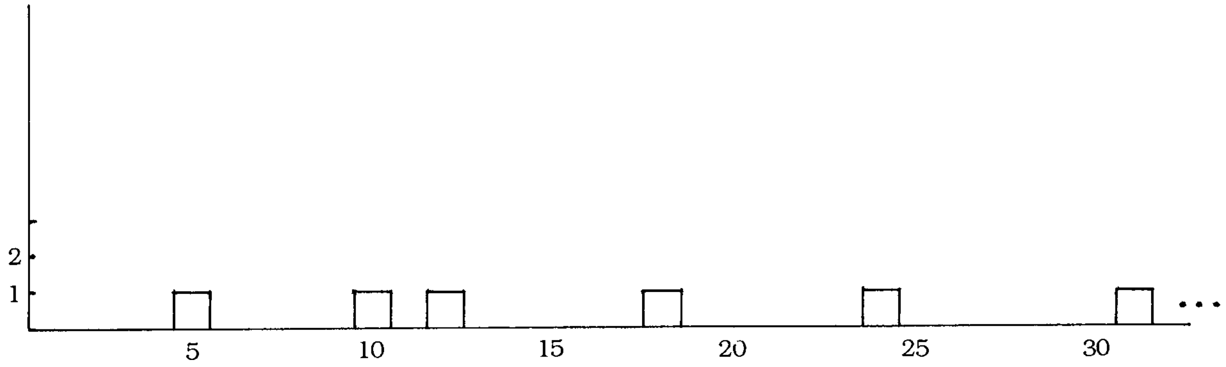

FIG. 8. A possible distribution of species abundances in a real community.

be species that are sparsely distributed in the sample area, in any case.

DISCUSSION

Mathematical ecology would benefit greatly from a general theory of sampling that establishes certain criteria which candidates for species-abundance distributions must fulfill. A theory along the lines suggested above would play this role, for starters, by ruling out truncation as a valid operation on a distribution which has been proposed to describe, in nontruncated form, abundances within communities.

What has been said here applies with nearly equal force to the negative binomial distribution (Pielou, 1975). More generally, this paper has already provided the tools for a uniform treatment of all proposed distributions of species abundance that are not ruled out by the foregoing consideration or which do not lead to nonintegrable expressions. For such distributions the Pielou transformation provides a bridge between the distribution of abundances in samples and those in the parent communities. In particular, if an inversion can be found for Eq. (3), expressing g as a function of f, then natural communities might be to some extent reconstructable from samples.

REFERENCES

Dewdney, A. K. 1997. A dynamical model of abundances in natural communities, *COENOSES* **12**(2–3), 67–76.

Dewdney, A. K. 1998. A dynamical model of communities and a new species-abundance distribution, *The Biological Bulletin*, submitted for publication.

Feller, W. 1957. "An Introduction to Probability Theory and Its Applications," Wiley, New York.

Fisher, R. A., Corbet, A. S., and Williams, C. B. 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.

Froberg, C.-E. 1985. "Numerical Mathematics: Theory and Computer Applications," Benjamin/Cummings, Menlo Park, CA.

Magurran, A. E. 1988. "Ecological Diversity and Its Measurement," Princeton Univ. Press, Princeton, NJ.

Mandelbrot, B. B. 1977. "Fractals, Fun, Chance and Dimension," Freeman, New York.

May, R. M. 1975. Patterns of species abundance and diversity, *in* "Ecology and Evolution of Communities" (M. L. Colby and J. M. Diamond, Eds.), pp. 81–120, Harvard Univ. Press, Cambridge, MA.

Pielou, E. C. 1969. "An Introduction to Mathematical Ecology," Wiley, New York.

Pielou, E. C. 1975. "Ecological Diversity," Wiley, New York.

Preston, E. F. 1948. The commonness, and rarity, of species, *Ecology* **29**, 254–283.

Preston, E. F. 1962. The canonical distribution of commonness and rarity, *Ecology* **43**, 185–215 and 410–432.

Williams, C. B. 1964. "Patterns in the Balance of Nature and Related Problems in Quantitative Ecology," Academic Press, London.

Zipf, G. K. 1965. "Human Behavior and the Principle of Least Effort," Hafner, New York.