Chapter 3. Sampling in practice and in theory

A major focus of this book, the relationship between "samples" and "communities" in terms of their abundances, must now be put on a precise footing by defining both terms, the first in theory, the second in practice. In theory, a *community* C is simply a finite collection (set) of disjoint finite sets called *species*. Each species consists of individuals, but these are not distinguished.

$$C = \{S1, S2, \ldots Sn\}$$

In practice, there are many ways to define a community, depending on the sometimes peculiar nature of the organisms being studied. In most practical applications the following definition will serve:

In a field setting, a *community* is the set of all organisms that, at a particular point in time, inhabit a defined volume (or area) of the biosphere and belong to a particular "grouping" G of species. The time element is important as it enables us to visualize a living community as frozen in time, as though we could wander through that still landscape and sample to our heart's content. In reality, of course, the community is dynamic and, to allow this possibility, we could speak of a community through time, a kind of three-dimensional movie, if one likes. In many instances, the field biologist, having sampled over a relatively short period of time, might think of the community that was sampled as a once-fixed entity that, since the sample was taken, has undergone some population fluctuations, as well as the addition or deletion of a species or two. It is then a slightly different community.

The grouping G referred to in the previous paragraph is simply a very general conceptual slot into which we might insert a taxonomic criterion for membership, as in a community consisting of all species of macrofungi in the phylum Basidiomycota. Or we might use habitat and trophism, as in all species of near-shore stream invertebrate saprotophs. In other words, although "community" is often defined as a taxonomic group, other criteria are possible without having (so far) compromised the methods explained here or the theory behind those methods. I will not attempt a definition of species, since differing species concepts used in the biosurvey literature seem to make no difference to the appearance of the logistic-J distribution in the metastudy described in Chapter 1. The reader may deploy the concept that best suits.

The connection between the theoretical and practical definitions just given is that a natural community, as defined operationally, automatically implies the set structure of the theoretical definition. The structure is largely invisible, of course, and only to be discerned, dimly enough, through samples taken of it.

If the sample is "random" enough, the biologist may be fairly confident that a species that is common in the sample is also common in the community. By the same token, a species that

occurs just once in a sample will tend to belong to smaller populations in the community. In most cases many of the lowest-abundance species will not show up in a sample at all. The notion of randomness will be discussed in the next section

An *unbiased* sample is one in which no population of the community is favoured by the sampling method. In other words, if a particular unbiased method tends to sample the ratio r of one species, it will tend to sample the ratio r of all the species. Thus if r = 0.1, the sampling method will pick up approximately 10 percent of the individuals in each species in the community (give or take the usual statistical fluctuations). But if the method consistently samples ten percent of one species and one percent of another, the method is biased toward the first species. Such samples will over-represent the first species or under-represent the second one -- or both.

All discussion of samples in this book assumes unbiased methods. Most biologists, particularly those who sample animal communities, tend to be sensitive to this issue and to take compensatory steps to eliminate bias from the sampling method. For example if one is sampling a community of butterflies, some species may be drawn preferentially to a particular species of plant. Sampling patches of such plants exclusively will cause the species in question to be over-represented in the sample. Sweeping a large, defined area with the greatest variety of plant types will certainly help to eliminate that particular source of bias -- as long as the sweeping trails are randomly selected (and recorded for subsequent resampling, if necessary).

The ratio r mentioned above is called the *sampling intensity*. Broadly defined, a random (unbiased) sample of a community consisting of N individuals (all species) will have size rN. Of course, in any one sample, there will be some variation -- statistical in nature -- in the tendency for a species of abundance N_i to show up rN_i times in the sample. But over a number of repeated samples, it will appear close to rNi times, on average. I am aware that some authors refer to the intensity of a sampling process in relative terms, as would be the case where the field biologist who took three samples in a particular area would be sampling more intensively than one who took only two, other things being equal. If necessary, one could call the version defined here the “absolute intensity”, but for the purposes of this book, I will let the present term stand.

Sampling intensity is a very important parameter in any sampling scheme. As will be shown in Chapter 5, no richness estimation technique can work reliably in the absence of knowledge about r. If one doesn’t know the intensity of one’s sample, no method exists -- or can exist -- that will reliably and accurately predict the number of species in the community being studied. Indeed, the accuracy of any such method will depend critically on the accuracy of one’s estimate of r. Various methods for estimating r are described in Section 3.2.

Finally, statisticians distinguish two major types of sample. In the standard statistical model, balls are drawn at random from an urn - colored balls if one likes. If each ball thus withdrawn is replaced, the operation is called *sampling with replacement*. If not, it is called *sampling without*

replacement. We will see that modern sampling techniques, as deployed over the five kingdoms of life (six, if one counts Archaea), come in both flavours. In the context of the richness estimation methods to be explored in Chapter 5, the distinction between sampling with and without replacement is relatively unimportant when sample intensity is small because the removal of individuals from the general population (in reality or in effect) hardly changes selection probabilities. The distinction becomes more important in Chapter 9 where the problem of accumulation curves is explored. The accumulation curve for sampling with replacement has a horizontal asymptote at "infinity", the curve for sampling without replacement terminates abruptly with a nonzero (positive) slope.

The remainder of this chapter is devoted to a variety of topics, including practical issues, such as estimating intensity and theoretical topics such as the general theory of sampling.

3.1 The variety of sampling activity

It is interesting to classify the methods that biologists have developed to assess natural communities by way of samples. The methods speak to the ingenuity of field biologists in studying natural environments in a manner that is appropriate for each kind of organism.

The most common forms of sampling with replacement involve plants and some animals such as birds or fish. A botanist carrying out a survey of oldfield plants, for example, may lay out several quadrats at random and record every plant contained within each quadrat without picking any of them. (Not removing a plant at all is equivalent to sampling-with-replacement, although an occasional voucher sample may be taken.) A point-count of birds is carried out by the zoologist walking a specified distance, then stopping for a specified period of time and noting all birds calling or visible within a given distance. Of course, in this case, there is no actual collecting done, but the technique amounts to the same thing. The birds are still there when the biologist leaves the area. A somewhat better example is the method of sampling bats by catching them in a mist net when they are flying. A specimen can be extracted from the net, identified, recorded, and then released. There is always the problem of overcounting in these techniques, although with bats the problem can be solved by marking the animal with a temporary stigma before releasing it. Similar remarks apply to some forms of sampling fish. Most fish-trapping techniques, from seine nets to kick-samples, to minnow traps to drag nets can be deployed in this manner. Electrofishing, however, may kill or injure the fish brought to the surface (stunned) by powerful electric currents). Small mammals such as moles, shrews, voles, (non-jumping) mice, and small mustelids may be captured in pitfall traps, again to be counted and released. In all the foregoing cases involving vertebrates, the group being sampled is usually well known and an expert field biologist can make an identification rapidly enough to release the animal before any harm is done.

Fungi occupy a special place in this summary of techniques because, at least with macrofungi, only the fruiting body is normally collected, leaving the organism with its mycelial network more or less intact. Microfungi are typically part of soil samples or samples of other substrates. They

may be identified through culturing methods or identified through DNA sequence analysis of substrate samples.

Sampling without replacement is most common in arthropod surveys. No other group of organisms has more collecting techniques applied to it and almost all involve sampling without replacement, "sacrificing" the animal, if you will. Arthropods may be captured in light traps, pitfall traps, malaise traps, Berlese funnels, emergence traps, aspiration bottles, pan traps, and by sweeping, fogging, hand-collecting, beating, observing, and so on. We note here that different collecting techniques carry a natural emphasis on different habitats. Night-flying insects may be drawn to a light trap, but insects that are not flying at the time will not appear there. Indeed, moths, for example, may be drawn preferentially to a light trap, some strongly, some only weakly. Such preferences undoubtedly introduce bias into the sample so taken. Ground (Staphylinid) beetles are natural victims for pan traps and pitfall traps but will not usually be collected by sweeping bushes with a net.

With the smallest organisms, namely protists and bacteria, a sample of soil or water is collected and transported back to the laboratory for examination. If not killed by staining or other chemical treatments, the organisms may be flushed into the sink once they have been examined, but they are never "replaced" in situ.

When populations are large and samples are relatively small (as in the foregoing paragraph), it makes little difference to the final assessment whether the sample was taken with or without replacement. Removing five individuals from a population of 1000 at random has very little effect on the balance of probabilities for a subsequent observation.

3.2 Estimating sample intensity

Readers will recall that the sample intensity r is simply the ratio of individuals in the sample to those in the community being sampled. In the case of plant surveys, r is easily estimated. Assuming that the field biologist has a defined polygon within which a survey will be taken, it is relatively easy to calculate r. For example, if the biologist records every plant in each of ten randomly placed one-metre square plots within the polygon, the intensity will be

$$r = 10/A,$$

where A is the area of the polygon in square metres. Mycologists collecting samples have roughly the same advantage as botanists when it comes to calculating r.

In the animal kingdom, the sampling process is bedeviled by the tendency of subjects to wander off or fly away, to be counted more than once or not at all. However, even here some groups are easier than others. For example, a north temperate waste field in late July might present a sea of wildflowers such as goldenrod, aster, wild carrot, etc. with clumps or patches of species mixed

uniform-randomly throughout the field. The flowers are attended by several orders of insect, especially hymenoptera. The fact that bees, wasps, and ants are constantly moving from plant to plant may not matter, provided that the pattern of movement prevails over the whole field. Carrying out an intensive count within a number of randomly placed quadrats may give one a good estimate of the density d (individuals per square metre). This would lead to a reasonable estimate of the total size N of the hymenopteran meadow community and the intensity of the sample will be

$$r = n/N,$$

where n is the size of one's sample. Variations on this technique may be applicable to other groups, the areal consideration being one component and the estimate of total population being the other. In general the technique calls for intensive sampling (census would be a better word) within a limited area or areas, then extrapolating the result to the area under study. A major conclusion of the work presented in this book is that one cannot hope to make accurate estimates of species richness in any sampling milieu without having a reasonably accurate estimate of r in hand. (See The Isotropic Principle in Section 6.1.) The problem presented by varying sample intensities is also discussed in Section 9.4 in a chapter called Open Problems.

3.3 How samples have been used: calculating biodiversity

In the biosurvey literature one frequently finds authors who wish to go beyond the mere presentation of the abundance data they have so painfully gleaned from nature. They wish to say something meaningful about the sample, usually in terms of its "biodiversity." But to do so, they face a bewildering choice of "biodiversities" to choose from.

Although the term "biodiversity" has been much used in recent decades, it turns out to have no generally accepted definition. Instead, it has many. They illustrate once again the confusion that pervades theoretical ecology, as explained in Section 1.2. Although these concepts of biodiversity are defined for sample data, they are frequently interpreted as descriptions of the communities from which the samples came. In the following we consider a community (or sample of it) with m species having abundances given by the integers $a_1, a_2, a_3, . \ a_m$. Some of the following biodiversity measures (indices) use abundances in relative form $p_1, p_2, p_3, \ldots, p_m$, where $p_i = a_i/N$, N being the total number of individuals in the community (or sample). The maximum abundance *max* is defined as the largest integer in the set $\{p_1, p_2, p_3, \ldots, p_m\}$

Simpson Index	$B_1 = \sum p_i^2$	(Simpson 1949)
Shannon Index	$B_2 = - p_i \ln(p_i)$	(Shannon 1949)
Shannon-E* Index	$B_3 = - B_2/\ln(max)$	(Pielou 1969)
Brillouin Index	$B_4 = (\ln(N!) - \sum \ln(a_i!))/N$	(Pielou 1969)
Margelef Index	$B_5 = (R-1)/\ln(N)$	(Clifford & Stephenson 1975)
Menhinick Index	$B_6 = R/VN$	(Whittaker 1977)

Berger-Parker Index $B_7 = max/N$ (Berger & Parker 1970)

There are still other measures of biodiversity, but they add nothing to the illustration. The first four indices use individual abundances, but the last three do not.

I will now compute values for each index as applied to three different sample shapes, representing three distinct types of "community." Figure 3.1 shows three histograms, each of which has 12 species and 36 individuals. The first histogram, albeit somewhat reduced for illustrative purposes, is typical of samples that commonly emerge from biosurveys. We will call it C1. The next two histograms are distinctly odd. In the univoltine distribution of Figure 3.1b all the species have the same abundance and in the extremal distribution of Figure 3.1c, one species has abundance 25, while the remainder all have populations of one individual each.

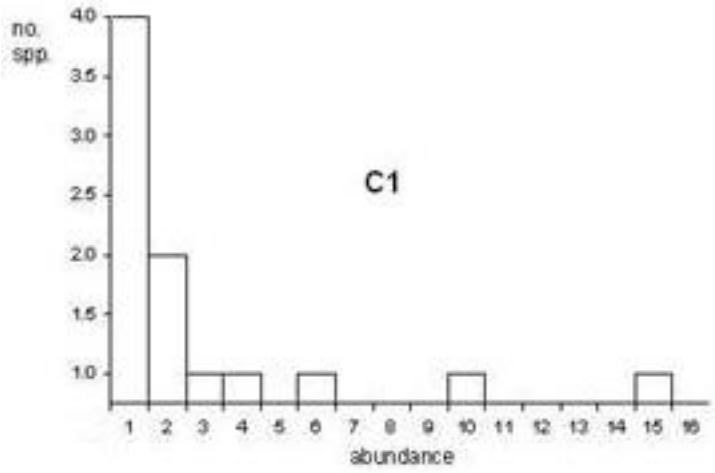

Figure 3.1.a A typical field histogram

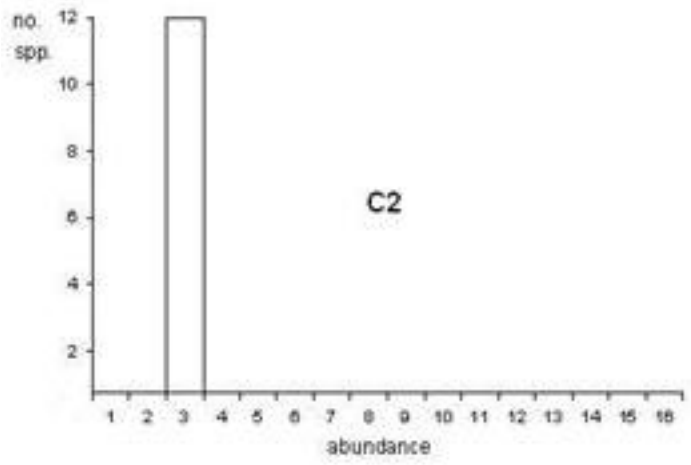

Figure 3.1b A univoltine distribution

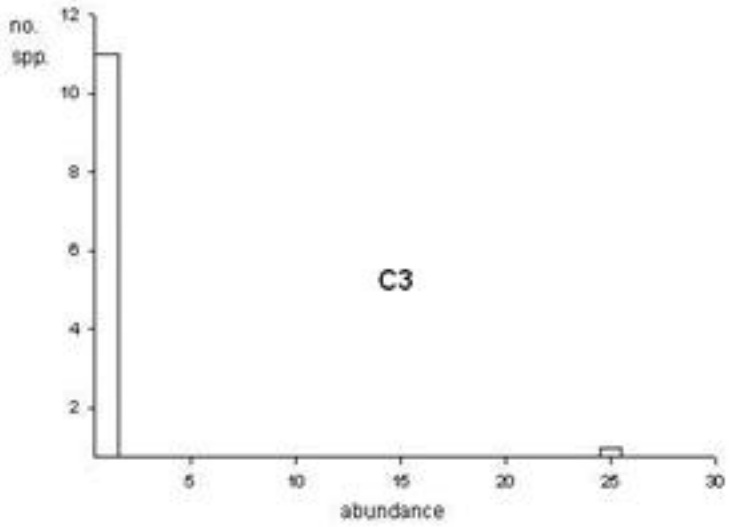

Figure 3.1c An extreme distribution

The table below displays the value taken on by each measure of biodiversity, for each of the three communities.

community:	**C1**	**C2**	**C3**
Simpson	0.116	0.084	**0.493**
Shannon	2.171	**2.485**	1.348
Shannon E	1.443	**29.81**	3.697
Brillouin	1.811	**2.062**	1.048
Berger-Parker	0.222	0.083	**0.694**
Margelef	3.070	3.070	3.070
Menhinick	2.000	2.000	2.000

Table 2.1. Biodiversity indices versus three different communities:
bold numbers are maxima

Although community C1 is most typical, in overall shape, of real communities, none of the first five indices give it their highest biodiversity assessment. Instead both Shannon indices and the Brillouin index give the highest score to what can only be described as a distinctly pathological community, one where all species have the same abundance. Since these three indices regard such a shape as in some sense the ideal, their use has limited value. The Berger-Parker and Simpson indices award the palm to the extremal community, again indicating a lack of contact with data. The remaining indices do not take individual abundances into account and so give the same score

to all three communities. Given that the intention behind the first five indices is to take not only the richness but the shape of a distribution into account, it must ba added that the project of developing a single numerical measure that embraces both aspects of a distribution is a losing proposition. It would be better to use at least two measures as a joint descriptor of "biodiversity." Otherwise, let this term stand as a synonym for species richness. As though the failings and confusion created by so many indices were already apparent, most biologists already follow the latter practice.

Gaston (1966) has summarised the problem of defining "biodiversity" as follows: "However, the abstract concept of biodiversity as the 'variety of life', expressed across a range of hierarchical scales, cannot be encapsulated in a single variable. The complexity in this sense is irreducible, and the search for the all-embracing measure of biodiversity, however desirable it might seem, will be a fruitless one." Gaston reminds us that "evenness" and "equability" were two of the goals of early biodiversity measures. Hurlburt (1971) has expressed similar views.

3.4 What is random?

In this section, I explore the concept of randomness with two applications in mind. The first concerns virtually all natural phenomena, including those that involve living organisms. The second aspect addresses randomness in samples. In order to conduct a successful sampling program, a field biologist must observe or collect individuals from the general (community) population "at random." Statistical methods can only be applied to such samples.

But what does the word "random" really mean? Most mathematical concepts are defined on the basis of the *presence* of certain properties. For example, a whole number is called *even* if it has 2 as a divisor. If a 2 is present among the factors of a number, the number is called even. But a number is called random if it is obtained in the *absence* of any systematic procedure. Of course, no single integer can ever be "random" by itself. Randomness is always defined in relation to a set of numbers, with an implicit, non-deterministic process of selecting individual numbers from the set -- over and over again, as necessary.

The only rigorous definition of randomness in relation to computers (and therefore deterministic processes generally) employs just such programs. According to the only widely accepted definition of randomness (Chaitin 2001), a sequence of length n is random if the minimum length of computer program that produces it increases as a function of n. In this definition, the sequence is potentially infinite. As (minimum-length) programs are found to generate longer and longer stretches of the sequence, it is found that the lengths of such programs grow in proportion to the lengths of the sequence.

This definition has little practical value, as the implicit test (writing all possible computer programs that produce a given sequence, then selecting the simplest of these) is far too complicated even to contemplate. There is a simpler definition of 'random' which, although not

as general as the foregoing, has immediate practical application. The base set from which numbers are drawn in this context is the binary set, {0, 1}.

A procedure for selecting items from this set is *effectively k-random* if all sequences of length k have an equal probability of appearing in the (larger) sequence generated by the procedure. The following little observation tells us that effective randomness is a hierarchical property. The proof will be found in Appendix A.3: If a procedure is k-random, then it is also (k-1)-random, for $k > 1$. This little theorem paves the way for a full definition of effective randomness: A procedure is *effectively random* if it is effectively k-random for all values of k that apply.

Any suitably long (truly) random sequence must be effectively random, but the converse is not true. However, the distinction makes little or no difference in practical applications.

In practice, the "random numbers" generated by a computer are produced by a deterministic program known as a *pseudorandom generator*, the sequences so produced being called *pseudorandom*. A traditional method of generating pseudorandom numbers is the *linear congruential generator*:

$$X_{n+1} = (aX_n + c) \bmod m$$

Starting with an initial or "seed" value X_0 for X, the program simply reiterates this basic equation, using the output of one iteration as input for the next one. The number m is called the *modulus*. Whatever the value that $(aX_n + c)$ might have, it is divided by m and the remainder taken as X_{n+1}. The constants a and c are called the *multiplier* and the *increment*, respectively.

For example, if $m = 8$, $X_0 = 5$, $a = 3$, $c = 11$, we have

$$X_{n+1} = (3X_n + 11) \bmod 8$$

and the sequence produced will be 5, 3, 4, 7, 0, 3, 4, 7, 0, etc. Clearly, this sequence repeats itself very soon. With a much larger modulus, this problem is overcome. Linear congruential generators are not much used these days, but the pseudorandom number generators of choice are not much more complicated. The random number generators in general use are effectively k-random for at least low values of k. They are presumably suitable for the experiments described in this monograph since they are suitable for commercial simulation (a multimillion dollar industry) as well as virtually all scientific simulation.

Whatever the status of (true) randomness in nature, effective randomness is very common, in the author's opinion. The factors affecting the fall of a seed that ultimately germinates cannot be known or predicted, in general, and this amounts, in itself to a thumbnail definition of randomness. What brought the rabbit to forage on a particular patch of vegetation this evening, as

opposed to a dozen other equally good ones? Again, the factors affecting the rabbit's decision are myriad and, for the most part, unknowable -- and unpredictable.

.

One of the most influential sources of randomness in nature is the weather. Given that weather systems tend to be chaotic, weather parameters tend to vary unpredictably over time in any given area or region. Although weather systems appear to have a Lorenz attractor (Lorenz 1963) at their dynamical core, much simpler chaotic systems yield the same random effects. For example, one can even use the logistic map (May 1976, Verhulst 1938) to generate pseudo-random numbers:

$$X_{n+1} = \lambda X_n(1 - X_n) \qquad \text{(i)}$$

The name "logistic map" requires an explanation. The word "map" (or "mapping") refers to an alternate name used by mathematicians for a function. The word "logistic" has the same meaning as it does in the name of our distribution, the Logistic-J, both systems having a finite population limit. In a minimal version of the multispecies logistic (MSL) system, the same function drives the process: if two species A and B have populations n and N-n, respectively, where N is the total number of individuals in the system. The probability that the population of A will increase at the next iteration must be

$$p(A) = n(N - n)/N^2,$$

since for each of the n ways of choosing an individual from A, there are (N - n) ways of choosing an individual from B. Replacing the notation n/N byX, one has,

$$p(A) = X(1 - X),$$

making the relationship clear. Unlike the MSL system, however, the logistic map leapfrogs over a myriad of incremental changes, whereas the MSL system fluctuates over a much finer time scale and with rather different results, in general.

Verhulst and May both studied the logistic map as a potential source of insight into the behaviour of populations. Equation (i) governs the direction and extent of population changes over tme: If the the populations N_1 and N_2 of two organisms sum to a fixed number N (the logistic limit) the populations may be expressed as ratios $X = N_1/N$ and $Y = N_2/N$. Equation (i) uses subscript to indicate successive values of the variable X and it uses as well the obvious relation, $Y = (1 - X)$.

In order for the equation to always produce values that lie in the interval [0, 1], we require that

$$0 \leq \lambda \leq 4,$$

since the function takes its maximum value, 0.25λ, when X = 0.5. The parameter λ could be called the *fecundity factor*, as it strongly influences the rate at which either population may grow. At low values of λ the variable X quickly converges to a specific number and remains there. At higher values of λ, the variable X alternates between two fixed numbers, then four, then eight, and so on, as λ is increased. The period-doubling behavior of the logistic map continues right up to a value of approximately 3.57 for λ, where chaos sets in. The variable X bounces around inside the interval [0, 1] with no seeming rhyme or reason. This feature of the equation is responsible for its popularity as an early population model in Ecology. Its appeal lay in its seeming realism, at least yielding behavior that was just as unpredictable as that of real populations.

A computer program that emulates the logistic map was used to generate 200 numbers in the interval [0. 1], of which a sample of 10 numbers are listed here by way of an example:

0.7681 0.6769 0.8310 0.5335 0.9457 0.1951 0.5966 0.9156 0.2971 0.7935 . . .

The numbers were produced by the logistic system with the parameter λ set to 3.8. In order to illustrate the presence of effective randomness in such numbers in a binary setting, each number has been replaced by the parity (even or odd) of the sum of its digits: 1 1 0 0 0 1 1 0 1 1 . . .

subsequence	frequency
00	14
01	13
10	10
11	13

Table 3.2. frequencies of two-bit subsequences in the random sequence

Performing a spectral analysis of the resulting bits, the sequence produced 51 0-bits and 49 1-bits, making it fairly 1-random, while the spectrum of consecutive pairs produced the frequencies shown in Table 2.2, again well within what might be expected from a 2-random source. Before going on to examine higher orders of effective randomness from such a source, the sequence of 200 “chaotic” numbers would have to be doubled, then doubled again, as the analysis proceeded. Most tests for randomness, from the poker test to the runs test, (Law and Kelton, 1999) are essentially specialized versions of the spectral test

3.5 Computer simulation of sampling

It is hard to see how anyone carrying out research in theoretical ecology, at least in the area of species abundance distributions, can expect to understand the sampling process without the

appropriate computational tools. The most important such tool is obviously one that simulates the sampling process itself. As a destroyer of hypotheses, it has no equal. The author has, on many occasions, seen a favorite hypothesis contradicted by simple computer experiments. Such experiences, as I have pointed out earlier, are the life blood of any genuine scientific enterprise.

Although most field biologists will not be simulating the sampling process, they should understand the relationship between "real" sampling and simulated sampling. Ideally, they are the same, but in practice they may not be. The idealized sampling process described in this chapter is an exact match with standard statistics -- to within the powers of pseudorandom number generators. In the context of the analysis of natural communities the samples taken must be random enough to ensure that a meaningful sampling intensity figure can be calculated and applied. The tools described here are indispensable, however, for those who study sampling theory in an ecological context.

Suppose we have in hand a distribution of N individuals from many species. This could be considered as a community, a portion thereof, or even a sample of a community. A random sample of N individuals completely ignores which species they may belong to. After the fact, one may construct a standard histogram, as described in the next chapter, and expect to see the community distribution reflected to some degree in the histogram.

In what follows, all samples will be with replacement, i.e., not removing any of the items of the sample from the original population (community) of N individuals. In other words, some individuals may be sampled more than once, in effect, even as others are not sampled at all. This process is readily simulated by a computer program that uses a random number generator.

In the context of simulated sampling, we will define a *sample of intensity r* as a random sample (with replacement) of rN individuals from the original population of N individuals. In the case where r = 1, we will call the resulting sample a *perturbation* of the distribution. (but see Sections 1.6.1 and 7.3.1 for more complete definitions.) A perturbation of a distribution may also be obtained by varying the entries in each abundance category according to the binomial distribution, a discrete version of the normal distribution.

3.5.1 A sample simulation algorithm

To simulate a direct sampling process is easy by computer. It takes an abundance distribution as input, either theoretical in the form of an expression, or empirical in the form of a list or biodiversity array,

$$(a_1, a_2, a_3, \ldots a_R),$$

where a_i represents the abundance of the ith species and R is the number of species. The

theoretical expression just mentioned can readily be converted into a biodiversity array, as used in the following algorithm. Random individuals are selected by generating a random number, then counting through all the species in the biodiversity array until that number is arrived at. Thus if the number 27 is chosen at random and the first five abundances in the array are

5, 9, 2, 8, 5, 5, 15,. . . ,

one would count through the array entries by adding up the abundances on the way. When the entry where the sum first equals or exceeds 27 is reached, the corresponding species is the one that individual #27 belongs to. Thus we have

5+9+2+8+5 = 29,

so the 27th individual belongs to the fifth species in the list. The algorithm that appears below uses two arrays, an array of species abundances that has the same structure as the biodiversity array and an array of species counts, both initialized to zero.

1. Input biodiversity array, along with values for N and r.
2. Set all species counts to 0 in the abundance array.
3. Let n be the integer {rN}. (greatest integer not exceeding rN)
4. Repeat the following steps n times:
    4.1. Choose a random number j from the interval [1, N].
    4.2. Calculate the species s of j (as described above) in the biodiversity array.
    4.3. Increment the count for species s in the abundance array.
5. For each species,
    5.1 Look up its abundance k as calculated in loop 4.
    5.2 Add 1 to the species count array at abundance k.
5. Print or display the array for use.

The sample simulation technique just described involves sampling with replacement. To convert it into a program that samples without replacement, one simply adds the additional sub-step,

4.4. decrement the count for that species in the biodiversity array

One may obtain a perturbation of a distribution in this manner by setting r = 1, as we have already seen. The algorithm just listed is the simplest possible. It may be adapted to any programming language one happens to be familiar with, but it is essential to follow the structure of calculations given here so as to ensure that individuals are sampled, and not species.

3.6 The general theory of sampling

The most important thing to know about the sampling process from a theoretical point of view is

that it expresses what mathematicians call a transformation, in this case of a community distribution into an image set -- the sample. The transformation has the important property that it preserves abundance patterns in a special way. If the community abundances follow the distribution G, then so do the abundances within the sample, albeit with different parameter values. Strangely, there was no general sampling theorem in the literature before 2002 (Dewdney 2002). If the theorem had been published earlier, say in the 1930s or 40s when it could have (and should) have been, much of the confusion resulting from multiple abundance distribution proposals could have been avoided.

The Pielou transformation (named in honour of Elizabeth Pielou, an eco-theorist who used a similar tool) maps the abundances in a community with the distribution F into an "expected" sample with distribution F', where

$$F'(k) = \int_0^\infty (e^{-rx}(rx)^k/k!)F(x)dx, \qquad \text{(ii)}$$

the integral being taken from 0 to ∞. The transformation is based on a statistically exact formula, the hypergeometric distribution, which involves a ratio of factorials that are somewhat cumbersome to work with mathematically or to program, for that matter.

However, the hypergeometric distribution is very closely approximated by the Poisson distribution (Hays and Winkler 1971), as it appears in formula (ii). One could argue that "exact ecology" should not involve approximations, however reverting to the hypergeometric distribution is always possible and, if need be, one could determine the approximation error for specific cases. Thus exactitude, although slightly compromised, is always within reach.

From a purely mathematical point of view (sampling theory aside for the moment), one may ask what form the function F' has if we set $F(x) = x^n$, the simplest form of polynomial, in the indefinite form of the integral:

$$F'(k) = \int (e^{-rx}(rx)^k/k!)x^n dx,$$

$$= r^{-n}\int (e^{-rx}(rx)^{k+n}/k!)dx,$$

$$= r^{-n}(k+n)!/k!\int (e^{-rx}(rx)^{k+n}/(k+n)!)dx$$

Evaluated between its limits, the integral equals unity, leaving

$$F'(k) = r^{-n}(k+n)!/k!$$

If the expression (k+n)! is multiplied out, one obtains a new polynomial in k of degree n. For example, if $F(x) = x^2$, we have (k+2)(k+1)

$$F'(k) = \quad = r^{-2}(k^2 + 3k + 2)$$

In this particular case, one may describe the effect of the transformation as follows. The function x^2 is an upward-opening parabola with its apex at the origin (0, 0), while F'(k) is an upward-opening parabola with its apex at the point (- 1.5, - 0.25). Moreover, the factor r^2, being less than unity, has the effect of reducing the height of the parabola. Summarizing the net effect, the Pielou transformation shifts the initial parabola 1.5 units to the left and flattens it in the process.

The transformation of a general polynomial of degree m may be regarded as the transformation of a sum of terms x^n, the result of which is a sum of polynomials of degree m. In turn, the sum of such polynomials of degree m is again a polynomial of degree m. The latter polynomial is a translated (horizontally) and a compressed (vertically) version of the former. Clearly, the Pielou transformation may operate on any continuous function, whether a distribution or not.

The general argument has two parts: The first part hinges on the recognition that the Pielou transform operates on distributions in which the objects drawn are individuals, rather than species. The function F must therefore have numbers of individuals in the ordinate position, unlike any distribution based on what I have called the "standard axis" in which abundances play this role. The simplest distribution to have this property is the rank abundance diagram. (See Section 4.3 for a fuller treatment.) This is simply a histogram in which the abscissa consists of integers 1, 2, 3, and so on, which denote the order of abundance. The height of the first column is thus the number of individuals in the most abundant species. The height of the second column is the next lasrgest population, and so on.

Each theoretical distribution G that is based on the standard axis may be inverted mathematically into an equivalent distribution F in which the roles of species and abundances are interchanged. The formula F may not have a convenient form but it will represent the initial distribution uniquely since inversion is a 1-1 operation. Moreover, if the starting function is smooth and continuous, the inverted form F will also have that property. Discrete functions like the log-series are readily replaced with matching continuous counterparts.

The sedond part of the argument proceeds by invoking the Weierstrass Uniform Approximation Theorem. (Hobson 1950) Any continuous function F can be approximated over its domain to an arbitrary precision by a polynomial expression, according to the Weierstrass uniform approximation theorem. One may therefore replace any distribution F(x) by a polynomial P(x) such that the inequality

$$|F(x) - P(x)| < \nu$$

holds over the domain of F for an arbitrarily small quantity ν. If we apply the Pielou transformation to the function P, we obtain a polynomial P' that can be made (by choosing ν small enough) to approximate F' to any desired degree of precision, as the following inequality makes clear:

$$\int (e^{-rx}(rx)^k/k!)|(F(x) - P(x)|dx \leq \nu\int (e^{-rx}(rx)^k/k!)dx \qquad \text{(iii)}$$

$$= \nu,$$

The integral on the right hand side of inequality (iii) is the area under the Poisson density function, namely unity. The degree of approximation of F(x) by P(x) is therefore inherited by F'(x) and P'(x). The inheritance of shape by P'(x) from P(x) therefore implies an inheritance by F'(x) from F(x). In the limit the approximation error is zero.

It now follows that for each value of k the following two integrals can be made arbitrarily close:

$$F'(k) = \int (e^{-rx}(rx)^k/k!)F(x)dx \quad \text{and} \quad P'(k) = \int (e^{-rx}(rx)^k/k!)P(x)dx$$

The functions F and F' therefore have the same general form, differing only in the values of their (common) parameters.

Finally, when the two functions are inverted once again, the same statement may be made about G and G'.

The chief application of the sampling theorem to date has been to imply that the "veil line" proposed by Preston (1948) is incorrect; the lognormal distribution, as portrayed on our standard species/abundance axes, would be mapped into another lognormal distribution and not a truncated one. In this particular case, the use of the a logarithmic transformation of the abundance axis serves to keep the lognormal concept alive. Analysed on standard axes, I doubt that it would survive, so to speak, especially in view of the observation in the next section. How would the logistic-j distribution survive in the logarithmic milieu?

3.6.1 effect of a log transformation on the logistic-J distribution

It may be asked what happens if one subjects the theoretical form of the logistic-J distribution to a logarithmic transformation of the abundance axis, as used in connection with the lognormal distribution. (See Section 4.2) The answer came as a surprise to me. (Dewdney 1998) A truncated, unimodal curve emerges that in many cases will appear bell-shaped, albeit semingly

truncated on the left..

To put the answer on a solid footing, the principal tool will be the general integral of the logistic-J distribution over the arbitrary interval [a, b]: Here, R' will represent the area under the curve which yields the number of species having abundances in this interval.

$$R' = Rc\int_{a}^{b}(1/(x+\varepsilon) - \delta)$$

$$= Rc(\ln((b+\varepsilon)/(c+\varepsilon)) - \delta(b-a))$$

Values for the integral may be calculated for the subintervals [0, 1], [1, 3], [3, 7], and so on, according to the scheme of Preston, but this time applied to a continuous function. It would yield essentially the same result if I used a discrete version of the distribution; the integral is simply a lot less work, with only one calculation per octave. Figure 3.2 displays the result when the log transformation is applied to the distribution LJ[2.0, 20.0] x 50.

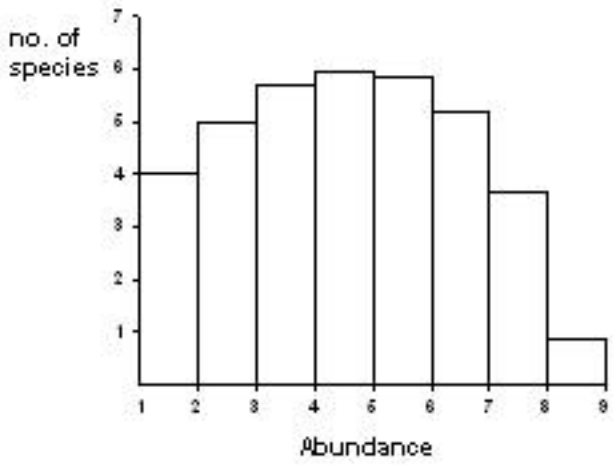

Figure 3.2. log transform applied to abundance axis of the logistic-J distribution

According to logistic-J theory, samples subjected to this treatment would resemble a perturbed version of the histogram of Figure 3.2. The species that go missing from a sample are "veiled" by a sigmoidal curve when viewed on standard axes (Dewdney 1998) and not by the veil "line" of Preston (1948). Researchers, such as Hubbell (2001), who use this representation run the risk of mistaking the logistic-J distribution for the lognormal. Indeed, Gaston (2005) has pointed out that the shape that results from logarithmic axis compression cannot be Gaussian (i.e., normal). The example in Figure 3.2 illustrates the dangers inherent in a representation that destroys information.