THE USE OF INFORMATION THEORY IN THE STUDY OF THE DIVERSITY OF BIOLOGICAL POPULATIONS

E. C. PIELOU
STATISTICAL RESEARCH SERVICE,
CANADA DEPARTMENT OF AGRICULTURE, OTTAWA

1. Introduction

Any natural assemblage of animals or plants usually contains several species of organisms. It may therefore be described as diverse. Only in the unlikely event that all the organisms in a collection belonged to the same species could one say that its diversity was zero.

Diversity is thus a characteristic of biological collections. Whether the object of study be a natural community of plants, a collection of insects caught in a light trap, the microarthropods in soil samples, a population of breeding pairs of forest birds, or the plankton organisms in a sample of sea water, it will almost always exhibit diversity. A biologist will therefore wish to assign some numerical value to this property of the collection he is studying.

Various methods of measuring diversity have been used in the past, the simplest being merely to count the number of species present. More precise measures take account of the fact that diversity has two quite distinct aspects. Thus, besides knowing the number of species in a collection, it is also necessary to consider how the individual organisms are apportioned among them. For a given number of species, a collection in which the species are fairly evenly represented has high diversity; whereas, if the bulk of the collection is made up of only a few of the species, while the remaining species are poorly represented, the diversity is lower.

Before describing in detail a way in which diversity may be measured, it is worth while considering why the diversity of populations is of interest to theoretical biologists. The diversities of numerous populations have been measured in various ways and facts such as the following have emerged: tropical communities are more diverse than those of high latitudes [1]; the communities of continental land masses are more diverse than those of isolated oceanic islands [1]; well established communities that have been undisturbed for long periods of time are more diverse than immature ones; communities of short lived organisms that show great seasonal variation in numbers (such as Lepidoptera in temperate latitudes) have high diversity in midsummer and low diversity in spring and fall [2]; sea floor animals living in shallow water form more diverse communities than do those at greater depths [2]; stable communities—those in which the numbers of the various species remain comparatively constant—are more diverse than unstable ones, in which great fluctuations in species abundance commonly occur [3].

Results such as these suggest some tentative generalizations. It appears that high diversities are characteristic of communities that are mature and stable, and that occupy "hospitable" habitats. Conversely, communities that are immature, unstable or exposed to rigorous environments are usually of low diversity [4]. The interconnectedness of the concepts of maturity, stability and diversity, and their dependence on environment, are of central importance to students of the evolution of biological communities. It is not possible to go further into these matters here; they have been fully discussed in [1], [5], [6], and [7].

2. Information content as a measure of diversity

To measure the diversity of a collection, it has been suggested [5] that its information content be used. For a collection of N individuals belonging to s species with N_i individuals in the ith species ($i = 1, 2, \cdots, s$; $\sum_{i=1}^{s} N_i = N$), the total information content (or diversity) is given by Brillouin's formula

$$B = \log \frac{N!}{N_1!N_2!\cdots N_s!} \tag{2.1}$$

units of information. The argument for using an expression of the form of B to denote the amount of information in a message is as follows [8]. Suppose a message is to be composed using a total of N symbols (letters of the alphabet or other code symbols) of which N_1 are of the first kind, N_2 of the second kind, $\cdots$ and N_s of the sth kind ($\sum_i N_i = N$). The number of possible messages that could be formed with these ingredients is obviously $N!/(N_1!N_2!\cdots N_s!)$. A recipient of the message who knows in advance only what symbols it will be composed of, but not what the particular message is, realizes that $N!/(N_1!N_2!\cdots N_s!)$ possibilities exist, all equiprobable, and this is the amount of his uncertainty before receiving the message. Once the message has been received and decoded this uncertainty is dispelled, and an equal quantity of information gained. The reason for using the logarithm of the number of possibilities rather than the number itself, will become clear below.

There is an obvious analogy between a biological collection consisting of various numbers of different species of organisms, and a coded message consisting of various numbers of different kinds of symbols. In identifying the members of a collection, one by one, and assigning each to its correct species, the actions of a biologist are formally identical with those of a man observing, one after another, the symbols in a message. There is therefore a property of biological collections analogous to that property of a message known as its information content. In the biological context, this property may be regarded as diversity.

It is easily seen that B has the properties we desire in a measure of diversity. For given N and s, B is a maximum when the individuals are divided among the species as evenly as possible. That is, when $N = N/s$ for all i, if N is a multiple of s; or otherwise, denoting the integral part of N/s by $[N/s]$ and putting $N = s[N/s] + r$, when $s - r$ of the species have $[N/s]$ individuals each and the remaining r species have $[N/s] + 1$ individuals each. Moreover B is a minimum when one of the species contains $N - s + 1$ individuals and the remaining $s - 1$ species have only one individual each.

Since one will often wish to compare the diversities of collections of different sizes, it is more convenient to consider $H = B/N$, the diversity per individual rather than B, the total diversity, of a collection. If now, throughout the formula

$$H = B/N = (1/N)\left\{\log N! - \sum_{i=1}^{s} \log N_i!\right\} \tag{2.2}$$

we substitute Stirling's approximation to the logarithm of a factorial in the form $\log x! = x(\log x - 1)$, the resulting approximation to H is

$$H' = -\sum_{i=1}^{s} \frac{N_i}{N} \log \frac{N_i}{N}, \tag{2.3}$$

or

$$H' = -\sum_{i} p_i \log p_i, \tag{2.4}$$

where $p_i = N_i/N$ for $i = 1, 2, \cdots, s$, denotes the proportion of the collection belonging to the ith species.

Whether to use H as in (2.2) or H' as in (2.3) when calculating the diversity of a biological collection will be discussed in the next section. First it is worth remarking on the connection between them.

The formula for H' is immediately recognizable as Shannon's [9] formula for the information per symbol in a message composed of s kinds of discrete symbols whose probabilities of occurrence are $p_1, p_2, \cdots, p_s$. But the formula for H' was not originally derived as a limiting form of H when all the N_i increase. Instead, it was arrived at as follows. Consider a complete system of s events, $X_1, X_2, \cdots, X_s$ with probabilities $p_1, p_2, \cdots, p_s$. As a measure of the information content (or uncertainty) of this system, we require a function of the probabilities, $H'(p_1, \cdots, p_s)$ say, which will satisfy the following conditions:

(i) H' must be continuous with respect to all its arguments;

(ii) H' must take its largest value when $p_i = 1/s$ for all i;

(iii) The addition of any number of impossible events, $X_{s+1}, X_{s+2}, \cdots$ for which $p_i = 0$, $i > s$, must not alter the information of the system. That is we must have

$$H'(p_1, \cdots, p_s, 0, \cdots, 0) = H'(p_1, \cdots, p_s). \tag{2.5}$$

Further, consider a second, dependent system of t events, $Y_1, Y_2, \cdots, Y_t$ and let the conditional probability of the occurrence of Y_j given that X_i has occurred be q_{ij}. Finally, consider the new complete system of the st joint events X_iY_j

for which the probabilities are $\pi_{ij} = p_i q_{ij}$ $(i = 1, 2, \cdots, s; j = 1, 2, \cdots, t)$. We require that the information content of this joint system, namely

$$H'(\pi_{11}, \pi_{12}, \cdots, \pi_{st})$$

should be given by

$$H'(p_1, \cdots, p_s) + \sum_i p_i H'(q_{i1}, \cdots, q_{it}). \tag{2.6}$$

It has been proved [10] that the only function satisfying these conditions is

$$H'(p_1, \cdots, p_s) = -C \sum_i p_i \log p_i \tag{2.7}$$

where C is a positive constant.

Besides the fact that Brillouin's $H(= B/N)$ tends to Shannon's H' as the N_i increase without limit, there is another connection between these formulae [11]. As before, consider an s species collection with N_i individuals in the ith species, with $\sum_i N_i = N$, and let its total diversity be B_0 as calculated from (2.1). Suppose now that we remove from the collection a single individual chosen at random. The probability that the chosen individual belongs to the ith species is thus N_i/N. Denote the expected diversity of the remaining population, of size $N - 1$, by $E(B_1)$. Then the expected decrease in diversity resulting from the removal is

$$\begin{aligned} B_0 - E(B_1) &= \log \frac{N!}{N_1! \cdots N_s!} - \sum_i \frac{N_i}{N} \log \frac{(N-1)!}{N_1! \cdots (N_i - 1)! \cdots N_s!} \\ &= -\sum_i \frac{N_i}{N} \log \frac{N_i}{N}. \end{aligned} \tag{2.8}$$

In words, the expected reduction in *total* diversity (as measured by Brillouin's B) is identical with the initial diversity *per individual* (as measured by Shannon's H').

Further, the expected change in H (the diversity per individual as calculated from Brillouin's formula) is [12]

$$\begin{aligned} H_0 - E(H_1) &= \frac{B_0}{N} - \frac{E(B_1)}{N-1} \\ &= \frac{-1}{N(N-1)} \log \left\{ \frac{N!}{N_1! \cdots N_s!} \left(\frac{N_1}{N}\right)^{N_1} \cdots \left(\frac{N_s}{N}\right)^{N_s} \right\}. \end{aligned} \tag{2.9}$$

The term in braces is equivalent to a multinomial probability, so is less than unity. It follows that $H_0 - E(H_1)$ is always positive. Thus we may conclude that the *expected* result of removing a randomly chosen individual from a collection will always be to decrease the collection's diversity per individual as measured by H.

3. The choice between H and H' as measures of diversity

Opinion among biologists appears to be divided on whether it is better to use H or H' to measure the diversity of biological collections. My own opinion is that when all the members of a collection can be identified and counted H is more suitable, for the following reasons.

(i) When measuring the information content (per symbol) of messages, H is the appropriate measure for particular messages whereas H' is defined only for average conditions in long messages [13]. A biologist using H' is therefore tacitly assuming that his collection may be regarded as a random sample from some much larger parent population and represents the average conditions in it. This is usually a dubious assumption and one that should not be made unless its reasonableness can be adequately demonstrated. If a particular collection is treated as an entity to be studied for its own sake, H is the proper measure of its diversity.

(ii) Even if one could regard one's collection as being representative of a large parent population, it is unsatisfactory to accept a single sample value of N_i/N as a reliable estimate of p_i, as is often done. This is especially true of the rare species in a collection, that is, those having low values of N_i. An attempt to increase the values of the N_i by enlarging the collection is no help, however, for when this is done one usually obtains not only more members of species already present but also, in ones and twos, individuals of species not previously represented. It is unusual for a collection not to contain at least some species with only one or two individuals [2]. Consequently, precise estimates of all the p_i, such as are needed to calculate H', are not easy to obtain (but see section 7).

(iii) It seems desirable that a measure of diversity should depend, not only on the number of species and their relative proportions, but also on the size of the collection. Compare, for instance, a two species collection with 5 members in each of the two species with another two species collection having 500 members in each. The number of possible arrangements of the individuals is clearly greater for the large collection than the small; this is the same as saying that the large collection has the greater diversity per individual. But if one were to put $p_1 = p_2 = 1/2$ for both collections, their H' values (= log 2 in this example) would be equal. However, H does depend on N, as well as on s and on the relative proportions; also, it is always less than H', the measure appropriate to conceptually infinite populations. To see this we note that

$$N(H' - H) = -\log\left\{\frac{N!}{N_1!\cdots N_s!}\left(\frac{N_1}{N}\right)^{N_1}\cdots\left(\frac{N_s}{N}\right)^{N_s}\right\}. \tag{3.1}$$

That is, $N(H' - H)$ is the negative of the logarithm of a probability and is therefore always positive.

It may be objected that H is an unsatisfactory measure of diversity when it is either impossible or undesirable to count the individuals in a collection. If a collection consists of a natural community of plants growing together in a defined area, it is impossible to count the individuals of those species that reproduce vegetatively; parents and offspring remain organically connected by rhizomes or stolons for an indefinite length of time and the very concept of "individual" becomes meaningless. And in a collection in which the members of the different species differ markedly in size it may be desirable to record the amount of a species by weight or volume rather than by number of individuals, even if they can be counted. In such cases it may seem natural to use H' instead of H but the objections to H' listed above still apply. It therefore seems best to allow N_i to denote the number of units of the ith species whatever they may be. For instance, in the plant population to be described in section 8, the quantity of plant material of each species was determined as the fresh weight in tenths of a gram of all above ground parts of the plants. These weights were then substituted for the N_i in the formula for H.

4. Units of diversity and the measurement of "evenness"

In the equations so far given the base of the logarithms has been deliberately left unspecified since the only effect of changing the base is to change the size of the units. Any desired base may be used; information theorists commonly use 2 and the units of information are then binary digits or bits. Using an arbitrary base, say n, one can define the units as "n-ary digits."

For instance, with an s species collection we might use s as the logarithmic base and measure the diversity by

$$H_{(s)} = \frac{1}{N} \log_s \frac{N!}{N_1! \cdots N_s!} \tag{4.1}$$

s-ary digits, where the bracketed subscript denotes the units.

Alternatively, if Shannon's formula were used, we should have

$$H'_{(s)} = -\sum_i p_i \log_s p_i \tag{4.2}$$

s-ary digits. The greatest possible value of $H'_{(s)}$, which would occur if the individuals were evenly distributed among the species, is thus

$$H'_{(s),\max} = -\sum_i \frac{1}{s} \log_s \frac{1}{s} = 1. \tag{4.3}$$

Clearly then, $H_{(s)}$ measures the diversity of a collection relative to the maximum attainable, for the same number of species, in a collection of unlimited size. In other words, $H_{(s)} = H/H'_{\max}$, the ratio of the actual diversity in arbitrary units (as calculated from Brillouin's formula) to the maximum possible diversity in the same units (as calculated from Shannon's formula). Margalef [5] has proposed that $H_{(s)}$ be used to measure what may be called the evenness component of diversity. It depends, however, on the size of the collection as well as on its evenness; a finite collection of maximum evenness, namely one in which all the N_i were equal, would still have $H_{(s)} < 1$ since, as already shown, $H_{\max} < H'_{\max}$. A better measure of evenness is provided by the ratio $H/H_{\max}$ [12], in which both terms are calculated from Brillouin's formula. Using the symbols of section 3,

$$H_{\max} = \frac{1}{N} \log \frac{N!}{\left\{\left[\frac{N}{s}\right]!\right\}^{s-t} \left\{\left(\left[\frac{N}{s}\right] + 1\right)!\right\}^{t}}. \tag{4.4}$$

In comparing several collections a biologist may compare separately the three properties: number of individuals N; number of species s; and evenness $H/H_{\max}$. Or he may prefer the single measure of diversity H using the same logarithmic base (that is, the same units) for all collections. Which course to adopt depends on the biological problem being investigated.

5. Hierarchical diversity

Although H is commonly called the "diversity" of a collection, it could be more exactly described as the specific diversity since it depends only on the numbers of individuals in each species. We now wish to take into account the hierarchical nature of biological classification. For simplicity consider only two taxonomic levels and let the individuals in a collection be sorted in two stages: at the first stage each individual is assigned to its genus; and at the second stage each genus is taken in turn and its individuals assigned to their species. Suppose that after the first sorting there are found to be g genera with N_i individuals in the ith genus with $i = 1, 2, \cdots, g$; $\sum_{i=1}^{g} N_i = N$. The generic diversity of the collection may then be defined as

$$H_G = \frac{1}{N} \log \frac{N!}{N_1! \cdots N_g!}. \tag{5.1}$$

After the second sorting the ith genus is found to contain s_i species with N_{ij} individuals in the jth of these species, where $j = 1, 2, \cdots, s_i$; $\sum_{j=1}^{s_i} N_{ij} = N_i$. The specific diversity within the ith genus is then

$$H_{S,i} = \frac{1}{N_i} \log \frac{N_i!}{N_{i1}! N_{i2}! \cdots N_{is_i}!}. \tag{5.2}$$

The specific diversity of the whole collection is

$$\begin{aligned} H &= \frac{1}{N} \log \frac{N!}{\prod_{i=1}^{s_1} N_{1i}! \prod_{i=1}^{s_2} N_{2i}! \cdots \prod_{i=1}^{s_g} N_{gi}!} \\ &= \frac{1}{N}\left[\log \frac{N!}{N_1! \cdots N_g!} + \sum_{i=1}^{g} \log \frac{N_i!}{N_{i1}! \cdots N_{is_i}!}\right] \\ &= H_G + \sum_{i=1}^{g} \frac{N_i}{N} H_{S,i}. \end{aligned} \tag{5.3}$$

Similarly, if three taxonomic levels are considered, family, genus and species, we may write

$$H = H_F + \sum_{i=1}^{f} \frac{N_i}{N} H_{G,i} + \sum_{i=1}^{f} \sum_{j=1}^{g_i} \frac{N_{ij}}{N} H_{S,ij}. \tag{5.4}$$

Here H_F is the familial diversity. $H_{G,i}$ is the generic diversity in the ith of the f families, which contains N_i individuals. And $H_{S,ij}$ is the specific diversity in the jth genus of the ith family; this genus contains N_{ij} individuals.

Analogous equations may be written for any number of taxonomic levels.

Consider again the case in which only two levels are recognized. An s species collection in which the individuals belonged to one or a few genera might well be thought of as less diverse than another s species collection with many (up to s) genera, although both could have the same value of H. A biologist wishing to allow for this might define a new measure of diversity, H^* say, by putting

$$H^* = \alpha H_G + \beta \sum_{i=1}^{g} \frac{N_i}{N} H_{S,i} \tag{5.5}$$

and then choosing weighting factors α and β with $\alpha > \beta$. There seem to be no "natural" values to assign to these constants and they would necessarily be arbitrary. Alternatively, instead of weighting and adding the components, H could be treated as a vector. The same approach could be used for any number of taxonomic levels.

Whether these more elaborate measures of diversity will prove useful in ecological research only experience can show. In the meantime, specific diversity as given by H is, in a sense, the most basic. If we assume that only conspecific individuals are interfertile, then grouping the individuals by species gives groups such that breeding is possible only within groups, and not between groups.

6. Pattern diversity in sessile populations

If a collection consists of motile or freely drifting organisms such as plankton in a water sample, it is impossible to do more than count the individuals in each species. But suppose we are investigating a community of sessile organisms such as forest trees; besides counting the individuals, we can also map their locations. It is then possible to study what may be called the community's pattern diversity [12] as well as its specific diversity. If the individuals tend to grow together in single species clumps we may say that the pattern diversity is low, or that the species are segregated from one another [14]. More thorough intermingling of the species would give higher pattern diversity.

Consider a population of size N and assume that the numbers of individuals N_i, with $i = 1, \cdots, s$, in each of the s species is known. Denote the specific diversity of any group of n neighboring individuals by $H(n)$. Then, on the hypothesis that the species are unsegregated, $H(n)$ has expectation

$$E\{H(n)\} = \sum \frac{n!}{r_1! \cdots r_s!} \frac{N_1^{(r_1)} \cdots N_s^{(r_s)}}{N^{(n)}} \cdot \frac{1}{n} \log \frac{n!}{r_1! \cdots r_s!}, \tag{6.1}$$

where the summation is over all r_i $(0 \leqq r_i \leqq n)$ for which $\sum r_i = n$.

Given an actual population, we may sample it at random, taking as sampling unit a group of n neighbors; and we may calculate $H(n)$, the diversity of each sampling unit and hence obtain $\overline{H}(n)$, the observed mean. If $\overline{H}(n)$ does not differ significantly from $E\{H(n)\}$, one may conclude that the species are unsegregated or randomly mingled. Otherwise, if $\overline{H}(n) < E\{H(n)\}$ significantly, it follows that there is segregation. It is convenient [12] to use the ratio $D = \overline{H}(n)/E\{H(n)\}$ as a measure of pattern diversity. Then for unsegregated populations $E(D) = 1$; low values of D correspond with low pattern diversity or a high degree of segregation; high values of D with high pattern diversity or a low degree of segregation. Estimation of the standard error of D is described in [12].

The use of this measure of pattern diversity unfortunately requires two arbitrary choices on the part of an investigator. These are:

(i) the method of defining "neighbors." Various possibilities will occur to ecologists and no single one is likely to be convenient in all circumstances. For example, one could take the n individuals closest to a random point; or the nearest individual to the center in each of n nonoverlapping sectors centered on a random point;

(ii) the value of n, the number of neighbors to have in each sampling unit.

Evaluation of $E\{H(n)\}$ and hence of D is impracticable with $n > 4$. It requires calculation of the probabilities of all terms of a multivariate hypergeometric series; and also of the H values corresponding to all partitions of n individuals into s groups of which some may be empty. Further, use of the multinomial distribution as an approximation to the hypergeometric will not usually be permissible since it is likely that at least some of the N_i will be as low as one or two.

The statistic D can be useful to ecologists, however, as may be shown by an example [12]. Six young populations of trees which had grown up on burned over land were being studied. The number of trees in each inevitably decreased as the populations aged since the young trees were too dense initially for all to survive to maturity. The numbers of species in the populations ranged from four to thirteen, and maps were available showing both the species and locations of all trees present at a certain time, and also those that survived throughout a period of time (which ranged from five to eleven years for the different populations). The answers to two questions were sought.

(i) Did natural thinning cause an increase or a decrease in the evenness component of diversity? A decrease would occur if the rare species were more likely to die than the common ones, so that in the course of time rare species became rarer and common ones commoner.

(ii) Did natural thinning cause a change in pattern diversity, and if so, in which direction? If deaths occurred chiefly within the single species clumps of young trees originally present, owing to competition among them, the result would be an increase in D (a decrease in segregation) with time. Conversely, if, out of the numerous tree seedlings that originally germinated on each area, each species survived only in sites to which it was adapted and died out elsewhere, the survivors would become increasingly well sorted or segregated and D would decrease.

To answer the first question the evenness, as measured by $H/H_{\max}$, was determined for each population at the beginning and end of the respective observation periods. In three of the populations there was an increase in evenness, and in the other three a decrease. There was thus no consistent tendency for the common species to survive at the expense of the rarer ones.

To answer the second question, D was estimated for all six populations at both dates. To do this $\overline{H}(n)$ was determined, with $n = 3$, by sampling each population at a number of random points and observing at each point the species of the nearest tree in each of three 120° sectors centered on the point. $E\{H(n)\}$ for $n = 3$ was calculated from (6.1). It was found that in all the five populations for which $D < 1$ significantly at the earlier dates (that is, those which were initially segregated), D increased with the lapse of time. This suggests that the natural thinning resulted from intraspecific competition within single species clumps of young trees.

In studying the evolution of a plant community it is clearly desirable to observe the changes that occur both in specific diversity, and in pattern diversity or segregation. In an unstable or maturing community, the number of species and their relative abundances can be expected to change, and also their spatial arrangement relative to one another. Attempts to interpret the way in which communities evolve should take both these changes into account.

7. Estimating the diversity of a large collection

As remarked in section 3, it is usually unsafe to assume that a particular collection constitutes a random sample from a larger population. For example, the insects caught in a light trap should not be thought of as representative of all insects within a certain radius, or even of all phototaxic insects within that radius. Differences among the species in flying power, and in opportunities to see the light, may exert a sorting effect. Also, the relative proportions of the different species that are active may vary during the hours that the trap is operating. A single catch should therefore be treated as a whole population and not as part of a larger one. Often, however, a single catch may contain several million insects and the collection is too big for all the individuals to be identified and counted. Then the population value of H cannot be determined and it becomes necessary to estimate the diversity of the collection from a sample. What is estimated is, of course, H', the average diversity per individual in the large population.

A method of doing this has been described by Good [15]. We assume it to be possible to mix the collection (now the "population") thoroughly and take from it a truly random sample. The sample need not contain representatives of all the species in the population; the fact that some rare ones may be missed does not affect the argument. We hypothesize that the (unknown) number of species is s, and that the population proportions of the species, in any order, are $p_1, p_2, \cdots, p_s$. We now identify and count the individuals in the sample and denote by ν_r the number of species represented by r individuals. Thus, if we have ranked the species in order of increasing abundance and, as before, have put N_i for the number of individuals in the ith species, we now have

$$(7.1)\qquad \begin{array}{lll} N_i = 1 & \text{for} & i = 1, 2, \cdots, \nu_1 \\ N_i = 2 & \text{for} & i = \nu_1 + 1, \nu_1 + 2, \cdots, \nu_1 + \nu_2 \\ \vdots & & \vdots \\ N_i = r & \text{for} & i = \sum_{j=1}^{r-1} \nu_j + 1, \sum_{j=1}^{r-1} \nu_j + 2, \cdots, \sum_{j=1}^{r} \nu_j. \\ \vdots & & \vdots \end{array}$$

Then $\sum_r r\nu_r = N$, the size of the sample.

Also, denote by q_r the unknown population proportion of an arbitrary species that is represented by r individuals in the sample. Good [15] then proves that the diversity of the population as measured by Shannon's formula, namely $H' = -\sum p_i \log_e p_i$ is given exactly by

$$(7.2)\qquad H' = \frac{1}{N}\sum_r rE(\nu_r)\left\{\frac{1}{r+1} + \frac{1}{r+2} + \cdots + \frac{1}{N-r} - \frac{d}{dr}\log_e E(\nu_r) - E[\log_e (1 - q_r)]\right\}.$$

We now wish to estimate H' from the data. First it is necessary to smooth the sequence $\nu_1, \nu_2, \cdots$ and replace it by the sequence $\nu'_1, \nu'_2, \cdots$. Ways of doing this are described in [15]. Using these smoothed values, Good then shows that as an estimator of population diversity we may write

$$(7.3)\qquad \tilde{H}' = \log_e N - \frac{1}{N}\sum_r r\nu'_r\left(1 + \frac{1}{2} + \cdots + \frac{1}{r} - \gamma + \frac{d}{dr}\log_e \nu'_r\right),$$

where $\gamma = 0.5772 \cdots$ is Euler's constant and the differentiation is performed graphically or numerically. Good does not give an estimator of Var $(\tilde{H}')$.

Three great advantages of estimating H' by this method are: (i) it does not require that the population proportion of the ith species be estimated by N_i/N; (ii) it is not necessary to know the number of species in the population; and (iii) no assumption is made as to any relationship between ν_r and r. Thus we need adopt no hypothesis such as Fisher's [16], that the expectations of ν_r form a logarithmic series, or Preston's [17], that they have a lognormal distribution.

Disadvantages of the method are: (i) it requires the existence of a parent population from which it is possible to draw a random sample of individuals; (ii) the sample must be large, since large values of ν_r (when r is small) are required to permit acceptable smoothing of the sequence $\nu_1, \nu_2, \cdots$.

8. Estimating the diversity of a plant community

Of all the collections or communities whose diversity a biologist may wish to measure, probably the hardest to deal with is a tract of herbaceous vegetation. Unless the vegetation is extremely sparse (as in a desert) or the area under study exceptionally small, it will usually be impossible to examine more than a small proportion of the total area. The area may be sampled with randomly placed quadrats and there is no difficulty in determining the diversity H of each separate quadrat. However, because of the patchiness of vegetation—the tendency for large single species clumps to occur—any one quadrat will contain only a small portion of the vegetation pattern and only a fraction of the total number of species. Therefore the contents of a single quadrat cannot be "representative" of the vegetation of the whole area and consequently $E(H)$, the expected within-quadrat diversity per individual will usually be considerably less than H'_{pop} the average diversity per individual in the whole population. To paraphrase Lloyd and Ghelardi [18], H is not a sample of something bigger; it is a measurement of something that exists on a local scale. How then can H'_{pop} be estimated? A possible method is as follows.

Suppose we take the quadrats in random order. Let $\mathcal{B}_1$ be the total diversity of the first. Add the data from the second quadrat to those of the first and call the total diversity of this combined pair of quadrats $\mathcal{B}_2$. For example, if the first quadrat contained species 1, 2, 3 and 4 in amounts N_{11}, N_{12}, N_{13} and N_{14} where $\sum_j N_{1j} = N_{1}.$ and the second quadrat contained species 3, 4 and 5 in amounts N_{23}, N_{24} and N_{25} ($\sum_j N_{2j} = N_{2}.$), then the combined pair of quadrats would have total diversity

$$\mathcal{B}_2 = \log \frac{(N_{1}. + N_{2}.)!}{N_{11}!N_{12}!(N_{13} + N_{23})!(N_{14} + N_{24})!N_{25}!}. \tag{8.1}$$

(The first digit in each subscript denotes the quadrat and the second the species.) Continue in this way and denote by $\mathcal{B}_k$ the total diversity of the pooled data from the first k quadrats.

Analogously, we may put

$$\mathcal{H}_k = \frac{\mathcal{B}_k}{\sum_{j=1}^{k} N_{j}.}, \qquad \mathcal{H}'_k = -\sum_i q_{ki} \log q_{ki}, \tag{8.2}$$

where $q_{ki} = (\sum_{j=1}^{k} N_{ji})/(\sum_{j=1}^{k} N_{j}.)$. That is, $\mathcal{H}_k$ and $\mathcal{H}'_k$ are the diversities per individual of the pooled contents of the first k quadrats as given by Brillouin's and Shannon's equations respectively.

If sampling were continued without replacement until the whole population was exhausted the final $\mathcal{B}_k$ would, of course, be identical with B_{pop}, the population value of the total diversity; correspondingly, $\mathcal{H}_k$ would be identical with H_{pop}. Suppose, however, it is feasible to sample only a small fraction of the population. If we examine the curve of $\mathcal{H}_k$ against k we should expect to find that $\mathcal{H}_k$ increases (not necessarily monotonically) at first; and that the curve then levels off (if the sample is large enough) becoming approximately horizontal for all k greater than some t. If this happens we may reasonably regard the first t quadrats taken together as providing a "representation" of the parent population. Then $\mathcal{H}_{t+1}, \mathcal{H}_{t+2}, \cdots, \mathcal{H}_{t+r}, \cdots$ are all estimates of H'_{pop} and we could take the final value as our chosen estimate. There is no way of determining its standard error however, since the $\mathcal{H}$'s are not independent.

To obtain an estimate whose standard error can be estimated we may proceed as follows. Consider $\mathcal{B}_k$, the total diversity of the pooled contents of the first k quadrats. As Baer [11] has shown (and see section 2), the removal of a single randomly chosen individual from any one of the quadrats causes an expected reduction in $\mathcal{B}_k$ of $\mathcal{H}'_k$. This is exact and is true for any k. For sufficiently large k, the discarding of a whole quadrat, say the kth (which contains $N_{k\cdot}$ individuals) will cause an expected reduction in total diversity of

$$E(\mathcal{B}_k - \mathcal{B}_{k-1}) \doteq N_{k\cdot}\mathcal{H}'_k. \tag{8.3}$$

The approximation will be closest when the contents of the discarded quadrat constitute a random sample from the contents of all k quadrats; it is probably still fairly good even when this condition is not met, provided k is large enough and the individual quadrats (and hence the values of $N_{k\cdot}$) are small enough. If, therefore, we put $(\mathcal{B}_k - \mathcal{B}_{k-1})/N_{k\cdot} = h_k$, say, for all $k > t$, we obtain a sequence $h_{t+1}, h_{t+2}, \cdots$ of independent random variables such that $E(h_{t+r}) \doteq \mathcal{H}'_{t+r}$. So if it is justifiable to assume that t or more random quadrats provide an adequate representation of the population, we may take $\bar{h}$ as an estimate of H'_{pop}, and also put $\text{Var}\,(\tilde{H}'_{\text{pop}}) = \text{Var}\,(\bar{h})$.

Example. It was desired to estimate the diversity of the ground vegetation in a 3000 sq m area of mixed woodland, using the data from a sample of 100 randomly thrown meter square quadrats. The number of species in all the quadrats taken together was 62, and the greatest number found in any single quadrat was 14. The contents of a quadrat were measured by cutting off all the plants in it at ground level, sorting them into species, and weighing the amount of plant material (fresh) of each species to the nearest tenth of a gram. These weights were then treated as the N_i values for calculating H for the quadrat. The mean of the 100 observed values of H was $\bar{H} = 1.227$ (using natural logarithms). The successive values of $\mathcal{H}_k$ were calculated as already described, and the way in which $\mathcal{H}_k$ varied with k is shown in figure 1. The final value, based on the total sample, was $\mathcal{H}_{100} = 3.214$. Taking $t = 70$ (a subjective choice), 30 values of h, namely $h_{71}, h_{72}, \cdots, h_{100}$, were obtained. It was found that $\tilde{H}'_{\text{pop}} = \bar{h} = 3.056$ and $\text{Var}\,(\bar{h}) = 0.026$, whence (assuming $\bar{h}$ is normally distributed) the 95 per cent confidence limits for H'_{pop} were $2.74 \leqq H'_{\text{pop}} \leqq 3.37$.

As had been expected, $\tilde{H}'_{\text{pop}}$ greatly exceeded $\bar{H}$. The ratio $E(H)/H_{\text{pop}}$ (of which $\bar{H}/\tilde{H}'_{\text{pop}}$ provides an estimate) can, indeed, be thought of as a measure of pattern diversity or of its converse, segregation. Yet another interpretation consists in regarding this ratio as measuring the "graininess" of the vegetation pattern; a coarse grained pattern is one in which the diversity within any single

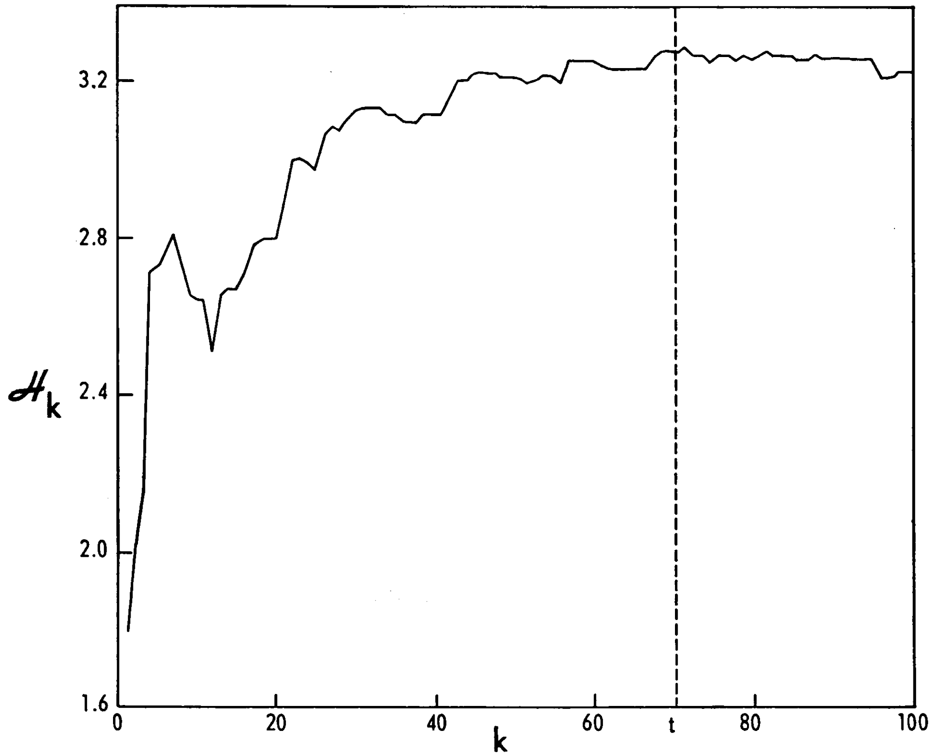

FIGURE 1

The relation between $\mathcal{H}_k$,
the diversity of the pooled contents of the first k quadrats, and k.

quadrat is low; the finer the grain, the greater the proportion of the pattern that any one quadrat can contain, and the more closely will $E(H)$ approach H_{pop}.

The magnitude of the ratio $\overline{H}/\tilde{H}'_{\text{pop}}$ will, of course, vary with the size of quadrat used since graininess is a relative property. A pattern can only be defined as fine grained (or coarse grained) relative to some arbitrarily chosen scale: a pattern that would be considered fine grained on a small scale (or when sampled with large quadrats) is coarse on a large scale (or when the quadrats are small).

9. Conclusion

Simpson [19] has written: "The aim of biology is to understand the structure, functioning, $\cdots$, and history of organisms and populations of organisms." One of these aims, the understanding of population structure, seems most likely to be gained by comparing the diversities of as many populations as possible, of various taxonomic groups, in various geographical regions and habitats, and at various times. This requires that the data gathered by many different people be comparable. So it is important that ecologists should agree on, and consistently use, a measure of diversity that is applicable to any population whatever. I believe that information content meets this requirement. It can either be determined or (for large populations) estimated for any collection that can be brought into the laboratory for sorting and counting, or for any community that can be delimited on the ground.

REFERENCES

[1] R. H. MacArthur, "Patterns of species diversity," *Biol. Rev.*, Vol. 40 (1965), pp. 510–533.

[2] C. B. Williams, *Patterns in the Balance of Nature*, New York, Academic Press, 1964.

[3] R. H. MacArthur, "Fluctuations of animal populations and a measure of community stability," *Ecology*, Vol. 36 (1955), pp. 533–536.

[4] J. H. Connell and E. Orias, "The ecological regulation of species diversity," *Amer. Nat.*, Vol. 98 (1964), pp. 399–413.

[5] D. R. Margalef, "Information theory in ecology," *Gen. Syst.*, Vol. 3 (1958), pp. 36–71.

[6] ———, "On certain unifying principles in ecology," *Amer. Nat.*, Vol. 97 (1963), pp. 357–374.

[7] R. H. Whittaker, "Dominance and diversity in land plant communities," *Science*, Vol. 147 (1965), pp. 250–260.

[8] L. Brillouin, *Science and Information Theory*, New York, Academic Press, 1962 (2nd ed.).

[9] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana, University of Illinois Press, 1963.

[10] A. I. Khinchin, *Mathematical Foundations of Information Theory*, New York, Dover, 1957.

[11] R. M. Baer, "Some general remarks on information theory and entropy," *Information Theory in Biology* (edited by H. Quastler), Urbana, University of Illinois Press, 1953.

[12] E. C. Pielou, "Species-diversity and pattern-diversity in the study of ecological succession," *J. Theoret. Biol.*, Vol. 10 (1966), pp. 370–383.

[13] S. Goldman, "Some fundamentals of information theory," *Information Theory in Biology* (edited by H. Quastler), Urbana, University of Illinois Press, 1953.

[14] E. C. Pielou, "Segregation and symmetry in two-species populations as studied by nearest neighbour relations," *J. Ecol.*, Vol. 49 (1961), pp. 255–269.

[15] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, Vol. 40 (1953), pp. 237–264.

[16] R. A. Fisher, A. S. Corbet, and C. B. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *J. Anim. Ecol.*, Vol. 12 (1943), pp. 42–58.

[17] F. W. Preston, "The commonness and rarity of species," *Ecology*, Vol. 29 (1948), pp. 254–283.

[18] M. Lloyd and R. J. Ghelardi, "A table for calculating the 'equitability' component of species diversity," *J. Anim. Ecol.*, Vol. 33 (1964), pp. 217–225.

[19] G. G. Simpson, *This View of Life*, New York, Harcourt, Brace, and World, 1964.