Field Studies 5 (1983) 825–832

SPECIES DIVERSITY: A COMMENT ON A PAPER BY W. B. YAPP

By M. B. USHER

Department of Biology, University of York, York YO1 5DD

ABSTRACT

Four measures of diversity are considered: the total number of species, S; the parameter α of the logarithmic series distribution; Shannon's index, H'; and Simpson's or Yule's index, D.

The total number of species is very sensitive to rare species inflating the number, and hence gives a poor description of diversity. The logarithmic series distribution, which is fitted to Yapp's (1979) data, gives a statistically good fit to the data but a systematic error indicates that rare species are under-estimated and common species over-estimated. Shannon's index is widely used, but gives results that are difficult to interpret. Variants of Simpson's index are discussed, and they are generally both easy to calculate and are immediately understandable (i.e. the probability that two individuals randomly selected will belong to different species).

The discussion of diversity and equitability indices concludes that Simpson's index probably has the most advantages for general use as it can be easily understood. Perhaps it is best to summarise diversity not by a single number but by quoting the number of species (S), a diversity index (D') and an equitability index (E_1).

INTRODUCTION

IN A RECENT part of *Field Studies,* Yapp (1979) compared four measures of diversity and applied them to his counts of woodland birds. He concluded that the Yule or Simpson Index was the most useful measure of diversity, though he did not make it clear why some measures of diversity should be rejected on scientific rather than intuitional grounds.

There are essentially two components of diversity—the variety of species, and the way in which the individuals are distributed among those species. These can be termed the "number of species" and the "equitability". Take as an example two samples each of 500 individuals. In one sample there are five species each represented by 100 individuals. The number of species is 5 and the equitability is maximised since all species have an equal number of individuals. In the second sample there are also five species, but this time the first species has 496 individuals, and the other four species have only one individual each. The number of species is the same as in the first sample, but the equitability is now minimised since there is no way in which these 500 individuals could be more unequally distributed amongst five species. Indices of diversity often attempt to incorporate both components of diversity, the number of species and the equitability, into a simple figure, or else they tend to neglect one or other of these components. This has given rise to some confusion, and to the often asked question "What does the index mean?" The aim of this note is to attempt to show how such a question can be answered or avoided.

The four measures of diversity that Yapp (1979) considered are listed below.

1. The total number of species recorded, S.
2. The parameter α in the logarithmic series distribution. This distribution is given by the sequence

$$\alpha x, \alpha x^2/2, \alpha x^3/3, \ldots, \alpha x^i/i, \ldots$$

where the term $\alpha x^i/i$ gives the number of species represented in the collection of S

species with exactly i individuals (e.g. the most uncommon species in the collection each have 1 individual, and it is predicted that there will be αx $(i = 1)$ of these species; the next most uncommon species will have 2 individuals each and one would expect $\alpha x^2/2$ $(i = 2)$ of these species, and so on). The sequence must, of course, sum to S, the total number of species. S is related to α by the expression for this summation,

$$S = \alpha \ln (1 + N/\alpha),$$

where N is the total number of individuals in the collection and ln y implies the natural logarithm of y (logarithm to base e). In the sequence quoted above, x is given by the expression, $N/(N + \alpha)$.

3. The index of diversity, associated with the name of C. E. Shannon, which is based on information statistics and which can be calculated as

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

where p_i is the proportion of the ith species in the collection, and the summation is over all of the species. The proportions are given by

$$p_i = n_i/N$$

where there are n_i individuals of the ith species. There are two aspects of this equation to note. First, p_i tends to zero for the most uncommon species, i.e. those species with only one individual in a large collection. Conversely, p_i has its maximum value for the commonest species in the collection. Second, although it is most usual to use numbers of individuals, one could also use some other measure of the abundance of a species in a community, for example the biomass, to calculate a diversity index.

4. An index of diversity associated with E. H. Simpson and sometimes referred to as Yule's Index. If p_i (as defined above) is the probability of selecting at random an individual of the ith species, then the index measures the probability of selecting a second individual that belongs to the same species. This probability is approximately p_i^2, and can be summed over all of the S species in the collection to give

$$D = \sum_{i=1}^{S} p_i^2.$$

More correctly, as the first individual selected reduced the total number of individuals of its species from n_i to $(n_i - 1)$ and the total number of individuals in the collection from N to $(N - 1)$, the index should strictly be written as

$$D = \sum_{i=1}^{S} \frac{n_i(n_i - 1)}{N(N - 1)}$$

since the probability of getting the first individual as species i is n_i/N and the probability of the second individual also being in species i is $(n_i - 1)/(N - 1)$. Both formulae give very similar values in practice, and hence the simple approximation is generally used. Since D is the probability that two individuals selected at random belong to the same species, then

$$D' = 1 - D$$

is the probability that the two individuals are of different species. As a community becomes more diverse, then the probability that two individuals are of the same species decreases, and hence D approaches zero and D' approaches 1. Many people prefer an index that becomes larger as diversity increases, and hence instead of using D it may be conceptually easier to use $d = 1/D$. The index of diversity, d, is not a probability since the smallest value that it can take is 1 (when $D = 1$, i.e. a community with no diversity) and it increases towards infinity as the community becomes more diverse and D approaches 0.

Criticisms of Yapp's paper

The Logarithmic Series Distribution

Yapp fitted the logarithmic series distribution to his birchwood counts of 2,225 individual birds belonging to 45 species. His data are reproduced in Table 1, although in fitting the distribution the value of α has been changed from 7.97 (given by Yapp) to 7.9885 so that the expression

$$S = \alpha \ln (1 + N/\alpha)$$

is virtually exactly 45, the number of species. Note also that Table 2 of Yapp's paper

Table 1. *Data for counts of birds in birch woodlands. The data in columns* n_i *and* O_i *are taken from Table 2 of Yapp (1979). The calculations for testing the goodness of fit of a logarithmic series distribution (column* E_j*) are described in the text.*

Number of individuals n_i	Number of individuals (grouped)	Group j	Observed number of species O_i(O_j grouped)	Estimated number of species (E_j grouped)	$O_j - E_j$	Contribution to χ^2
1	1	1	9	7.96	1.04	.1359
2	2	2	4	3.97	0.03	.0002
3	3	3	3	2.63	0.37	.0521
4 5 6	4–7	4	2 4 2 } 8	5.95	2.05	.7063
8 9 13 15	8–15	5	1 1 1 1 } 4	5.57	−1.57	.4425
16 22 24 30	16–31	6	2 2 1 1 } 6	5.23	0.77	.1134
44 45 55 58	32–63	7	1 1 1 1 } 4	4.76	−0.76	.1213
70 86 127	64–127	8	1 1 1 } 3	4.02	−1.02	.2588
151 175	128–255	9	1 1 } 2	2.89	−0.89	.2741
374 799	256+	10	1 1 } 2	2.02	−0.02	.0002
Total	—	—	45	45.00	0	2.1048

lists 2,225 individuals whereas in Table 4 of his paper it is given as 2,243 individuals—in both instances there are 45 species.

Yapp rounds the numbers derived from the logarithmic series distribution to the nearest whole number, and hence he indicates that no species may have more than 14 individuals. Of course such a treatment of a mathematical distribution is misleading since his actual data have long series of zeros, i.e. he found no species with between 375 and 798 individuals. Clearly the correct approach is to attempt some grouping of the distribution, as has been done here in Table 1. Any grouping is rather arbitrary, but the grouping used in the table is based on logarithmically increasing group sizes. The groups include all species with counts of 2^n to $(2^{n+1} - 1)$ individuals (the groups thus contain all those species with 4 to 7, 8 to 15, 16 to 31, 32 to 63, etc., individuals, each range inclusive).

With these grouped data, visual comparison of the fourth and fifth columns of Table 1 indicates that the logarithmic series distribution gives a reasonably good approximation to the actual distribution of counts. Closer inspection, by looking at the errors, i.e. $(O_j - E_j)$ in the sixth column of the table, indicates that these tend to be positive for the rarer species and negative for the commoner species. This means that the logarithmic series distribution is under-estimating the numbers of rarer species and over-estimating the numbers of the more abundant species.

Statistically, the correspondence between the observed and fitted distributions can be tested using a χ^2 goodness of fit test. This is given by

$$\chi^2 = \sum_{j=1}^{g} (O_j - E_j)^2/E_j$$

where there are g groups, O_j is the observed (or actual) number of species in the jth group, and E_j is the estimated number of species in that group using the prediction of the logarithmic series distribution. χ^2 has $(g-2)$ degrees of freedom. The difficulty of using the χ^2 test is that the expected values should not be too small: Steel and Torrie (1980, pp 529–530) indicate that no value of E_j should be less than 1. For the grouped data in Table 1 the value of χ^2 is 2.105, with 8 degrees of freedom, which is not statistically significant with P (probability) $\leqslant 0.05$. From the point of view of the χ^2 test, which does not take into consideration the signs of the $(O_j - E_j)$ deviations, the data give a satisfactory fit to the logarithmic series distribution.

In doing any testing between observed and expected distributions, it is important that their totals should be the same. In Table 1 it can be seen that the totals of both the fourth and fifth columns are 45.

Shannon's Index of Diversity

The information statistic, H', is easily, if tediously, calculated. For the data given in Table 1 it is calculated as follows. For those 9 species (fourth column, first row, in Table 1) which are represented by only one individual, the proportion of that species in the community, p_i, is $1/2225 = 0.0004494$, and $\ln 0.0004494 = -7.7076$ (natural logarithms). The contribution of each of these species to H' is

$$p_i \ln p_i = 0.0004494 \times -7.7076 = -0.003464,$$

or -0.03118 (-0.003464×9) for the group of 9 species. Similarly, for the 4 species with 2 individuals each (second row of Table 1), p_i is $2/2225 = 0.0008989$, and the contribution of each species to H' is

$$0.0008989 \ln 0.0008989 = -0.006305,$$

or -0.02522 for the group of 4 species. This process continues for all of the n_i values in Table 1 until one reaches the commonest species for which p_i is $799/2225 = 0.3591$, and its contribution to H' is

$$0.3591 \ln 0.3591 = -0.3678.$$

Adding all of these contributions together gives

$$H' = 2.359.$$

The calculations are shown in Table 2. The difficulty with this statistic is to understand its meaning. It is not in any way a probability, as implied in the editor's footnote at the bottom of p. 49 of Yapp's paper. Perhaps the best way to understand H' is to ask what the extreme values could be in a community with 2225 individuals belonging to 45 species. The least diverse community is the one that has 44 species represented by one individual each, and the 45th species with 2,181 individuals. H' in this hypothetical community has the value 0.172. The most diverse community is the one in which all of the species are equally common. This can be represented by a community of 20 species each with 50 individuals and 25 species each with 49 individuals. H' for this hypothetically diverse community is 3.807. It can be seen that the actual value of H' lies somewhere between these two extremes, but still the question can be asked as to what it actually means.

One thing is clear: the estimation of H' has nothing whatever to do with a sequence of prime numbers. Yapp misinterprets a paper on number theory, where integers are used, with the information statistic where he correctly introduces p_i in the calculation of H' as the probability of the ith event. Probabilities generally take non-integer values, and the two integer values that they can take, 0 and 1, would have no interest in a sequence of prime numbers. By introducing this digression into theorems of prime numbers, Yapp has certainly made the information statistic more difficult to understand in an ecological context. To view this statistic in a wider context, Hill (1973) shows that it is just one of a whole range of a possible continuum of indices.

Simpson's or Yule's Index

As noted in the Introduction, there is a certain amount of confusion as to what is a probability or a chance and what is not. Probabilities have bounds. A probability of zero, the smallest value that a probability can take, means that the event will never happen. In biological applications, generally there is a certain amount of rounding, and hence a probability written as zero in reality means a very small number approaching zero, and hence in words this would represent an event that was exceedingly unlikely to happen rather than an event that never happens. Similarly, the maximum value that a probability can take is 1, which strictly is saying that an event must always happen. In practice a probability of one means a probability that is approaching 1, and hence the event that we are considering is exceedingly unlikely not to happen.

Given these rules for what are probabilities, what can one make of the Yule Index given by Yapp? He says '. . . Yule's has a real meaning; it is a measure of the chance that, if two occurrences are taken at random, they will not be of the same species'. Because Yapp essentially uses d, as defined above, his values in his Tables 3, 4 and 5

Table 2. *Data derived from those listed in Table 1. The following symbols are used:* f_i *are the numbers of species with an observed number of* n_i *individuals, e.g. 9 species were represented by only 1 individual in the collection of 2225 individuals. The probability of an individual selected at random belonging to that species is* p_i, *i.e. 1/2225 in the first row. Data are generally given with 4 significant figures.*

n_i	p_i	f_i	$-p_i \ln p_i$	$-f_i \cdot p_i \ln p_i$	p_i^2	$f_i \cdot p_i^2$
1	.000449	9	.003464	.03118	+	.000002
2	.000899	4	.006305	.02522	+	.000003
3	.001348	3	.008911	.02673	.000002	.000005
4	.001798	2	.01136	.02273	.000003	.000006
5	.002247	4	.01370	.05481	.000005	.000020
6	.002697	2	.01595	.03191	.000007	.000015
8	.003596	1	.02024	.02024	.000013	.000013
9	.004045	1	.02229	.02229	.000016	.000016
13	.005843	1	.03005	.03005	.000034	.000034
15	.006742	1	.03370	.03370	.000045	.000045
16	.007191	2	.03549	.07097	.000052	.000103
22	.009888	2	.04565	.09129	.000098	.000196
24	.01079	1	.04886	.04886	.000116	.000116
30	.01343	1	.05806	.05806	.000182	.000182
44	.01977	1	.07758	.07758	.000391	.000391
45	.02022	1	.07889	.07889	.000409	.000409
55	.02472	1	.09147	.09147	.000611	.000611
58	.02607	1	.09507	.09507	.000680	.000680
70	.03146	1	.1088	.1088	.000990	.000990
86	.03865	1	.1257	.1257	.001494	.001494
127	.05708	1	.1634	.1634	.003258	.003258
151	.06787	1	.1826	.1826	.004606	.004606
175	.07865	1	.2000	.2000	.006186	.006186
374	.1681	1	.2997	.2997	.02825	.02825
799	.3591	1	.3678	.3678	.1290	.1290
Totals	—	45	—	2.359	—	0.177

all lie in the range 5.4 to 16.8, and hence they cannot be probabilities. Indeed his values are the inverse of the probability that two randomly selected individuals will be of the same species.

Using Yapp's birchwood data, as shown in Table 1, one obtains estimates of $D = 0.177$, $D' = 0.823$ and $d = 5.65$. The calculations are shown the first three and last two columns of Table 2.

CONCLUSIONS

How one measures diversity is very much a personal matter. A whole diversity of diversity measurements is listed by Southwood (1978, chapter 13), and it is very much a matter of personal preference which index is used. One must also consider whether one wants to emphasise the number of species, or the equitability, as these two components of diversity will interact with each other.

Yapp has correctly argued against the use of S, the total number of species, since this is dependent upon the area searched, or the size of the sample, as well as upon the time spent searching. Such relationships for plants on limestone pavements have been documented by Usher (1980). Hence, we can agree to eliminate that index at the start.

The logarithmic series distribution can also be eliminated. This is not for the reason given by Yapp, who failed to undertake the grouping needed to combine a lot of very small frequencies. The problem with Yapp's application of this distribution can be seen in his Table 2 where his "S actual" column totals to 45 whilst his "S calculated" column only sums to 28. The real reason why the logarithmic series distribution can be dropped is that it generally does not describe biological distributions very well. There is an extensive literature on the alternative lognormal distribution (see Preston, 1962, for a discussion of this distribution and some of its properties), which has been shown empirically to fit a large amount of biological data (see Whittaker, 1972). Just as α in the logarithmic series distribution can be used as an index of diversity, so can an index be derived from the lognormal distribution (Bulmer, 1974). It is likely that the lognormal diversity index will generally be better than α, since the distribution generally gives a better fit to biological data, but again it is difficult to interpret what the statistic actually means.

This leaves only the Yule and the Shannon indices from those listed by Yapp. The Simpson (Yule) index, in its D form, gives an estimate of the probability that two individuals selected at random from the community will be of the same species, and as such it is easily intelligible. It suffers from the disadvantage of decreasing as the diversity increases, and hence the statistic $D' = 1 - D$ should probably be used. It gives the probability that two individuals drawn at random from the community will belong to different species, and it increases from 0 (or near zero) in the least diverse communities to 1 (or near one) in the most diverse communities.

Although from practical considerations the Simpson index may be preferred to the Shannon index, they can both be used to estimate indices of equitability (May, 1974). Equitability indices tend to 1 for communities where the individuals are distributed evenly amongst all of the species, and they tend to 0 in a community with a highly non-uniform distribution of relative abundances (i.e. where one species is completely dominant and all the others are exceedingly rare). For the Simpson index, the equitability index can be written as

$$E_1 = 1/(DS) \text{ or } E_1 = 1/\{(1-D')S\} \text{ or } E_1 = d/S.$$

For Yapp's birchwood data it can be seen that

$$E_1 = 5.65/45 = 0.13 \text{ (or } 1/(0.177 \times 45) \text{ or } 1/\{(1 - 0.823) \times 45\}).$$

With the Shannon index there is also a simple form of equitability index, defined by

$$E_2 = e^{H'}/S \text{ or } E_2 = \exp(H')/S$$

where e is the base of natural logarithms (2.7183). Again, using Yapp's birchwood data, it can be seen that

$$E_2 = e^{2.359}/45 = 0.24.$$

These equitability indices tend to be closely correlated with each other if they are calculated for a lot of communities (see the data for termites in Usher, 1975), and hence although their numerical values differ it tends not to matter which of the two is used.

In conclusion, it seems best to use whatever index of diversity is the simplest to calculate and to understand. But diversity is a complex subject (Pielou, 1975, shows this only too clearly), and more understanding may be derived from studying the

data by means of tables or histograms or graphs to see just what has been collected. The index of diversity or the index of equitability is a summary statistic, summarising in a single number a large amount of information, and it is therefore not surprising that imperfections may be found in its summary. Simpson's index, in the form that gives a probability that two individuals drawn at random will be in different species (D'), probably gives the most readily intelligible summary, in the form of an index of diversity. Add to this the number of species, S, and the equitability index, E_1, and with the three numbers there is probably as useful a summary of diversity and its components as can be calculated simply.

ACKNOWLEDGEMENTS

The subject of diversity, and its measurement, was raised during a course on Conservation Evaluation at Malham Tarn Field Centre in August 1982: I should like to thank the participants of that course for discussing the subject and I hope that this note clarifies the concepts for others as well as ourselves. I should particularly like to thank Dr R. H. L. Disney, the warden of Malham, and Dr J. A. Fowler, a participant on the course, for their most helpful comments on drafts of this note; the obscurities that remain are mine, not theirs.

REFERENCES

BULMER, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30, 101–110.

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

MAY, R. M. (1974). General Introduction. In: *Ecological Stability*, ed. by M. B. Usher and M. H. Williamson, pp 1–14. Chapman and Hall, London.

PIELOU, E. C. (1975). *Ecological Diversity*. Wiley, New York, etc.

PRESTON, F. W. (1962). The canonical distribution of commonness and rarity. *Ecology*, 43, 185–215 and 410–432.

SOUTHWOOD, T. R. E. (1978). *Ecological Methods, with particular reference to the Study of Insect Populations*. Chapman and Hall, London.

STEEL, R. G. D. and TORRIE, J. H. (1980). *Principles and Procedures of Statistics. A Biometrical Approach, 2nd edn.* McGraw-Hill, Kogakusha, Tokyo, etc.

USHER, M. B. (1975). Studies on a wood-feeding termite community in Ghana, West Africa. *Biotropica*, 7, 217–233.

USHER, M. B. (1980). An assessment of conservation values within a large Site of Special Scientific Interest in North Yorkshire. *Field Studies*, 5, 323–348.

WHITTAKER, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, 21, 213–251.

YAPP, W. B. (1979). Specific diversity in woodland birds. *Field Studies*, 5, 45–58.

M. B. Usher: Species diversity: a reply to a paper by W. B. Yapp.

This paper is written with the author's usual clarity and precision. I think it makes a useful contribution to Field Studies and should be published. I do, however, think that the title should be "species diversity: a comment on a paper by W. B. Yapp", rather than "a reply to". One of Usher's strongest ripostes concerns Yapp's misinterpretation of a number theorem. But Yapp himself had already dismissed it as having any relevance to bird studies, so that is not really a matter of contention.

There are a couple of general points on which it would be interesting to have the author's comments.

1. Usher very rightly stresses the point that Simpson's Index has a meaning in terms of probability which is easy to understand. However, since most indexes get larger with increasing diversity it is consistent to invert it or to do something similar. That being the case, would it not be worth pointing out that we then have a measure of the community which is comparable with the variance of a set of measurements? Yule's and Simpson's Indexes are comparable with V and I for the measurements.

2. Instead of simply inverting Simpson's Index to get a "variance-type" index, why not use $-1n\ D$? This has the same limits as H for a given set of data, and also allows equitability to be calculated (as $-1n\ D/1n\ S$), if that is needed.