COMPUTER CALCULATION OF INFORMATION — THEORETIC MEASURES OF DIVERSITY

JERROLD H. ZAR
Zoology Department, University of Illinois, Champaign

ABSTRACT. — Computational methods are suggested which facilitate implementation of computer calculations of both the Shannon and Brillouin measures of information content.

The calculation of the information content of a collection of entities has been performed for several sorts of biological considerations (Quastler, 1953), and the use of information content as a measure of ecological diversity has been of special interest in recent years (Margalef, 1957; MacArthur, 1965; Pielou, 1966a, 1966c). An aggregation of items which is high in information content is high in uncertainty, entropy, and diversity, and the information content of ecological communities has also been considered in relation to stability (MacArthur, 1955; Leigh, 1965) and "maturity" (Margalef, 1963).

METHODS

The mean diversity, per individual (i), in a collection of N individuals belonging to s categories (e. g. species) may be calculated by the measure of Shannon (1948):

$$H' = -\sum_{i=1}^{s} p_i \log p_i \qquad (1)$$

or by that of Brillouin (1956):

$$H = \frac{1}{N!} \log \left(\frac{N!}{n_1!\, n_2! \ldots n_s!} \right) \qquad (2)$$

where n_i is the number of individuals in the ith category and p_i is n_i/N, the proportion of N which belong to category i.

These two measures are not identical. The relative merits of each are dependent upon the population sampling represented by the data (Pielou, 1966b, 1966c). Lloyd, Zar, and Karr (1968) have shown that given a set of data, the calculations of H and H' are equally simple if certain manipulations are performed. Thus the type of sampling one's data represent, and not the ease of calculation, should dictate which diversity measure to compute. For ease of calculation, Shannon's measure may be written as:

$$H_b' = \frac{c}{N} \left[N \log_a N - \sum_{i=1}^{s} n_i \log_a n_i \right] \qquad (3)$$

and the Brillouin formula may be written as:

$$H_b = \frac{c}{N} \left[\log_a N! - \sum_{i=1}^{s} \log_a n_i! \right] \qquad (4)$$

so that one needs only a table of $log\ n!$ and $n\ log\ n$ to calculate both H and H' with equal ease; such a table is available in base 10 (Lloyd, Zar, and Karr, 1968). The calculations may be performed using logarithms of any base, a, and the resultant H or H' may be expressed in any other base, b, by inserting the appropriate scale factor, c, in formulae (3) and (4). Bases commonly used are e, 2, and 10, where the unit of information is called the "nit", "bit", and "hartley', respectively (Table 1).

For workers performing these information-theoretic calculations routinely, a digital computer program offers great conservation of time and insurance against miscalculations. The efficiency of such computer programs is greatly enhanced by the use of formulae (3) and (4) rather than the parent formulae (1) and (2). Using formula (3) eliminates the need of computing the proportions $p_i = n_i/N$ a total of s times and then summing the s number of $p_i\ log\ p_i$ values. Formula (4) avoids the possibility of overflowing computer capacity when calculating factorials, for addition of logarithms is used for the

TABLE 1.—Scale factors (c) to convert from logarithmic base a to base b.

b	a		
	2	e	10
2.....	.....	1.442695	3.321928
e.....	0.693147	.....	2.302585
10.....	0.301030	0.434294	.....

necessary multiplication of the integers and factorials. The calculation of $log\ n!$ as $\sum_{m=2}^{n} log\ m$ is not without disadvantage, however, even on a computer. Regardless of the accuracy of each calculation of $log\ m$, there is some finite error associated with it. These errors, however small, are being summed, and generally result in a low approximation of $log\ n!$. Using eight significant figures throughout the calculations, the calculation of $log\ n!$ by the summing of logarithms may accumulate an error in the sixth significant figure well below $n = 50$. The use of "double precision" computer arithmetic serves to delay the appearance of the error, but it may do so at considerable cost in computer storage and time. The use of the logarithmic form of Stirling's formula for approximation of factorials:

$$\log n! = (n+\tfrac{1}{2}) \log n + \tfrac{1}{2} \log 2\pi - n \log e \qquad (5)$$

(Selby, 1965: 340) allows one to estimate the $log\ n!$ value desired in one algebraic computation. While this approximation is good only for large n, the related formula:

$$\log n! = (n+\tfrac{1}{2}) \log n + \tfrac{1}{2} \log 2\pi - \left(n - \frac{1}{12n} + \frac{1}{360n^3}\right) \log e \qquad (6)$$

(Fisher and Yates, 1963: 129) can estimate $log\ n!$, for n as small as 4, accurate to at least the 6th significant figure ($n = 1$ through $n = 2000$ have been tested). If natural logarithms are used, $log\ e = 1$ and the equation is somewhat simplified. It may also be noted that n need not be an integer to be used in this formula.

If the data are a large random sample, the standard error of H':

$$\text{S.E. } (H') = \sqrt{[\Sigma p_i \log^2 p_i - (H')^2]/N} \qquad (7)$$

(Basharin, 1959) may be calculated readily as:

$$\text{S.E. } (H') = \frac{c}{N} \sqrt{\Sigma n_i \log^2 n_i - (\Sigma n_i \log n_i)^2/2N} \qquad (8)$$

Also of interest is the determination of the maximum diversity possible with the given s and N. This may be calculated as:

$$H_{max}' = \log s \qquad (9)$$

or

$$H_{max} = \frac{c}{N} [\log N! - (s\text{-}r) \log D! - r \log (D+1)!] \qquad (10)$$

where D is the integer part of N/s and $r = N - s \cdot D$.

The evenness of the collection (Pielou, 1966c) is $J = H/H_{max}$ using the Brillouin measure, and $J' = H'/H_{max}'$ using Shannon's formula.

The minimum diversity possible is given as:

$$H_{min} = \frac{c}{N} \left[\log N! - \log Q! \right] \qquad (11)$$

or

$$H_{min}' = \frac{c}{N} \left[N \log N - Q \log Q \right] \qquad (12)$$

where $Q = N - (s - 1)$. The redundancy, R, is:

$$R = \frac{H_{max} - H}{H_{max} - H_{min}} \qquad (13)$$

and of course R' may be computed in an analogous fashion using Shannon type measures throughout. One may also arrive at an index of heterogeneity for the collection as

$$R_H = 1 - R \quad \text{or} \quad R_H' = 1 - R'.$$

A FORTRAN IV computer program has been prepared which calculates H,

H', S.E. (H'), H_{max}, H_{max}', H_{min}, and H_{min}' in bases e, 2, and 10, and also calculates J, J', R, and R'. User's options include the choice of computing the Shannon and/or Brillouin measures. A program listing, with instructions for use, is available on request.

Acknowledgments

These computer procedures were developed using the IBM 360 and IBM 7094/1401 facilities of the University of Illinois (partially supported by National Science Foundation grants, with time granted to the Zoology Department from Public Health funds of the University of Illinois Research Board). Many of the program features were suggested by James R. Karr.

Literature Cited

Basharin, G. P. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. Theory of Probability and its Applications 4:333-36. (Transl. from Teoriya Veroyatnose i ee Primeneniya.)

Brillouin, L. 1956. Science and information theory. Academic Press, New York.

Fisher, R. A. and F. Yates. 1963. Statistical tables for biological, agricultural and medical research. 6th ed. Hafner Publishing Co., New York.

Leigh, E. 1965. On the relation between productivity, biomass, diversity, and stability of a community. Proc. Nat. Acad. Sci. 53: 777-83.

Lloyd, M., J. H. Zar, and J. R. Karr. 1968. On the calculation of information-theoretical measures of diversity. Amer. Midland Natur. 79: 257-72.

MacArthur, R. 1955. Fluctuations of animal populations and a measure of community stability. Ecology 36: 533-36.

Margalef, R. 1957. Information theory in ecology. General Systems 3: 36-71. (Transl. from Mem. Real Acad. Sci. Artes, Barcelona 32: 373-449).

———— 1963. On certain unifying principles in ecology. Amer. Natur. 97: 357-74.

Pielou, E. C. 1966a. Species-diversity and pattern-diversity in the study of ecological succession. J. Theoret. Biol. 10: 370-83.

———— 1966b. Shannon's formula as a measure of specific diversity: its use and misuse. Amer. Natur. 100: 463-65.

———— 1966c. The measurement of diversity in different types of biological collections. J. Theoret. Biol. 13: 131-44.

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Tech. J. 27: 379-423, 623-56.

Silby, S. M. (ed.) 1965. Abridged mathematical tables. The Chemical Rubber Co., Cleveland.

Manuscript received November 4, 1967.