Shannon's Formula as a Measure of Specific Diversity: Its Use and Misuse
Author(s): E. C. Pielou
Source: *The American Naturalist*, Vol. 100, No. 914 (Sep. - Oct., 1966), pp. 463-465
Published by: The University of Chicago Press for The American Society of Naturalists
Stable URL: http://www.jstor.org/stable/2459245

LETTERS TO THE EDITORS

Correspondents alone are responsible for statements and opinions expressed. Letters are dated when received in the editorial office.

SHANNON'S FORMULA AS A MEASURE OF SPECIFIC DIVERSITY: ITS USE AND MISUSE.

Ecologists are making increasing use of "information content" to measure the diversity of many-species populations (Margalef, 1958; Hairston, 1959; MacArthur and MacArthur, 1961; MacArthur, 1964; Lloyd, 1964; Lloyd and Ghelardi, 1964; Pianka, 1966). Diversity is thus equated with the amount of uncertainty that exists regarding the species of an individual selected at random from a population. The more species there are and the more nearly even their representation, the greater the uncertainty and hence the greater the diversity. Information content, which is a measure of uncertainty, is therefore a reasonable measure of diversity. However, two different definitions of information content have been proposed and ecologists appear to be uncertain as to when each is appropriate.

Consider first a population all of whose members have been identified and counted. Let there be N individuals in s species with N_i individuals in the ith species ($i = 1, 2, \ldots, s$). Then, using the formula given by Brillouin (1960), the diversity (or information) per individual is

$$H = (1/N)\,(\log N! - \sum_{1}^{s} \log N_i!) \text{ information units.}$$

In using this formula we are determining, not estimating, the true population value of the diversity. Since all the N_i are known exactly, the problem of sampling error does not arise; it is zero.

Next, consider a population too large for its members to be counted. N and the N_i are then unknown, Brillouin's formula cannot be used, and we cannot determine the true population value of the diversity. Instead, it must be estimated from a sample. Let us postulate that the population is large enough to be regarded as effectively infinite, and that it contains s species in proportions $p_1, p_2, \ldots, p_s$. Then the population value of the diversity (or information) per individual is given by Shannon and Weaver's (1963) formula:

$$H' = -\sum_{1}^{s} p_i \log p_i \text{ information units.}$$

There are various ways of estimating H' and they are described elsewhere (Good, 1953; Miller and Madow, 1954; Basharin, 1959; Pielou, 1966a). Dif-

ferent methods are required depending on whether the number of species in the population is known (it may exceed the number found in a sample), and on whether one can draw a random sample from the population (this may not be feasible when a population of sessile organisms is arranged in a patchy manner). Some methods do not require the concurrent estimation of the p_i, whose population values are unknown.

The purpose of this communication is to point out the pitfalls of taking $-\sum_i (N_i/N) \log (N_i/N) \equiv H''$, say, as a measure of diversity. This formula is capable of two quite different interpretations and should never be used without a statement as to which is meant. The interpretations are:

(i) If the N_i used in the formula are population values, and if also for all i, N_i is sufficiently large for $N_i (\log N_i - \log e)$ to provide an acceptable approximation to $\log N_i!$; then $H = H''$ approximately, and H'' may be taken as an approximation to, not as an estimate of, the true population diversity, Brillouin's H. Although $H'' > H$ always, one may choose to treat the difference as negligible in order to save computing factorials. There is no sampling error, but only the error that results from using an approximate formula. The approximation is seldom necessary, however, since a table of $\log x!$ for $x = 1, (1) \ldots, 1000$ is available (Pearson and Hartley, 1954).

(ii) If the N_i are sample values, then H'' is the maximum likelihood estimator of the unknown population diversity H', obtained by taking the observed N_i/N as estimates of the unknown p_i. H'' is a biased estimator of H', however. Using natural logarithms and neglecting terms of order N^{-2}, its expected value is $H' - (s - 1)/2N$, where s is the number of species in the population (not the sample). To adjust for the bias, therefore, one must know s. Moreover H'', being an estimator, is subject to sampling error. An estimator of its large sample variance is given by

$$\hat{\text{var}}\,(H'') \doteq \frac{1}{N}\left\{\sum_1^s \frac{N_i}{N}\left(\log_e \frac{N_i}{N}\right)^2 - (H'')^2\right\}$$

(see Miller and Madow, 1954; Basharin, 1959; Pielou, 1966a).

It should be emphasized that H'' cannot be regarded as exactly equal to H' even when the population N_i are exactly known. To do so is to make two implicit assumptions: that the population at hand is a sample from some undefined, conceptually infinite "super-population," and that it is an exactly representative sample. This is because H' is strictly defined only for an infinite population (Khinchin, 1957); it can therefore never be exactly determined, but only estimated.

LITERATURE CITED

Basharin, G. P. 1959. On a statistical estimate for the entropy of a sequence of independent random variables, p. 333–336. *In* N. Artin [ed.], Theory of probability and its applications. Vol. IV. (Trans-

lation of Teoriya Veroyatnostei i ee Primeneniya.) Society for Industrial and Applied Mathematics, Philadelphia.
Brillouin, L. 1960. Science and Information Theory. 2nd ed. Academic Press, New York.
Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264.
Hairston, N. G. 1959. Species abundance and community organization. Ecology 40:404–416.
Khinchin, A. I. 1957. Mathematical foundations of information theory. Dover Publications, New York.
Lloyd, M. 1964. Weighing individuals by reproductive value in calculating species diversity. Amer. Natur. 98:190–192.
Lloyd, M., and R. J. Ghelardi. 1964. A table for calculating the "equitability" component of species diversity. J. Anim. Ecol. 33:217-225.
MacArthur, R. H. 1964. Environmental factors affecting bird species diversity. Amer. Natur. 98:387–397.
MacArthur, R. H., and J. W. MacArthur. 1961. On bird species diversity. Ecology 42:594–598.
Margalef, D. R. 1958. Information theory in ecology. Gen. Syst. 3:36–71.
Miller, G. A., and W. G. Madow. 1954. On the maximum likelihood estimate of the Shannon-Wiener measure of information. AFCRC-TR-54-75, Air Force Cambridge Research Center, Air Research and Development Command, Bolling Air Force Base, Washington, D.C.
Pearson, E. S., and H. O. Hartley. 1954. Biometrika tables for statisticians. Cambridge University Press.
Pianka, E. R. 1966. Latitudinal gradients in species diversity. Amer. Natur. 100:33–46.
Pielou, E. C. 1966a. The use of information theory in the study of the diversity of biological populations. Proceedings, Fifth Berkeley Symposium on Mathematical Statistics and Probability. (In press)
Shannon, C. E., and W. Weaver. 1963. The mathematical theory of communication. Univ. of Illinois Press, Urbana.

E. C. PIELOU

STATISTICAL RESEARCH SERVICE
CANADA DEPARTMENT OF AGRICULTURE
OTTAWA
May 9, 1966

THE EFFECT OF POPULATION DENSITY ON MORTALITY AND SEX RATIO IN THE MILKWEED BUG, *ONCOPELTUS*, AND THE COTTON STAINER, *DYSDERCUS* (HETEROPTERA)

In most animals that have been studied, females seem to live longer and to be more resistant to stress than males (cf. Lack, 1954, for review). There are two major theories as to why this is so. In the first, it is assumed that the homogametic sex, usually the female, is hardier because deleterious loci in the X-chromosome are not expressed. In the second, males are assumed