Measurement of Diversity

THE 'characteristic' defined by Yule[1] and the 'index of diversity' defined by Fisher[2] are two measures of the degree of concentration or diversity achieved when the individuals of a population are classified into groups. Both are defined as statistics to be calculated from sample data and not in terms of population constants. The index of diversity has so far been used chiefly with the logarithmic distribution. It cannot be used everywhere, as it does not always give values which are independent of sample size; it cannot do so, for example, when applied to an infinite population of individuals classified into a finite number of groups. Williams[3] has pointed out a relationship between the characteristic and the index of diversity when both are applied to a logarithmic distribution. The present purpose is to define and examine a measure of concentration in terms of population constants.

Consider an infinite population such that each individual belongs to one of Z groups, and let $\pi_1 \ldots \pi_Z$ ($\Sigma\pi = 1$) be the proportions of individuals in the various groups. Then λ defined as $\Sigma\pi^2$ is a measure of the concentration of the classification. It can take any value between $1/Z$ and 1, the former representing the smallest concentration or largest diversity possible with Z groups, and the latter complete concentration, all the individuals being in a single group. λ can be simply interpreted as the probability that two individuals chosen at random and independently from the population will be found to belong to the same group.

Now suppose a sample of N individuals to be chosen at random from a population of this kind, and let $n_1, n_2 \ldots n_Z$ ($\Sigma n = N$) be the numbers of individuals falling into the various groups. It is easily shown that $l = \dfrac{\Sigma n(n-1)}{N(N-1)}$ is an unbiased estimator of λ; this is almost obvious since $\frac{1}{2}N(N-1)$ is the number of pairs in the sample and $\frac{1}{2}\Sigma n(n-1)$ is the number of pairs drawn from the same group. l is also an unbiased estimate of λ when the sample-size varies, provided no samples of size 0 or 1 are included and that the probability of the sample $(n_1, n_2 \ldots n_Z)$ splits into these two factors:

$$P(n_1, n_2 \ldots n_Z) = P(N)\frac{N!}{n_1!\,n_2!\ldots}(\pi_1)^{n_1}(\pi_2)^{n_2}\ldots,$$

where $P(N)$ gives the probability distribution of the sample size, $2 \leqslant N \leqslant \infty$. This is true in particular when samples are obtained by the 'fixed-exposure' method common in biological work, N having then a Poisson distribution adjusted for the absence of the first two terms.

If repeated samples of size N are drawn from the same population, the values of l obtained will be distributed about λ with variance

$$\frac{4N(N-1)(N-2)\,\Sigma\pi^3 + 2N(N-1)\,\Sigma\pi^2 - 2N(N-1)\,(2N-3)\,(\Sigma\pi^2)^2}{[N(N-1)]^2};$$

or, if N be very large, approximately

$$\frac{4}{N}[\Sigma\pi^3 - (\Sigma\pi^2)^2].$$

The third and fourth cumulants of the distribution of l have also been calculated exactly. They indicate that as N increases, the distribution tends to normality except when $\lambda = 1/Z$; in that case the distribution of lNZ tends to that of χ^2 with $Z - 1$ degrees of freedom, but with its mean moved from $Z - 1$ to N.

The characteristic defined by Yule[1] is, in the notation used above, $1{,}000\ \Sigma n(n-1)/N^2$, which differs from l, the sample estimator of λ, only in having N instead of $N - 1$ in the denominator and in the scale factor of 1,000.

Now let us see what value λ takes for a population containing Z groups the frequencies of which are $\pi_i = w_i/\Sigma w$, where the w_i are chosen at random and independently from the Type III distribution

$$dF = \frac{1}{(k-1)!}\,e^{-w}\,w^{k-1}\,dw, \quad 0 \leqslant w \leqslant \infty.$$

This may be called a 'negative binomial population', since samples drawn from it by the 'fixed exposure' method will obey the negative binomial distribution. The value of λ appropriate to it is obtained by averaging $\Sigma w_i^2/(\Sigma w_i)^2$ over all sets $(w_1, w_2 \ldots w_Z)$ which can be drawn from the population of values of w. Thus

$$\lambda = \int_0^\infty \cdots \int_0^\infty \left[\frac{1}{(k-1)!}\right]^Z e^{-\Sigma w}\,[w_1 \ldots w_Z]^{k-1}\,\frac{\Sigma w_i^2}{(\Sigma w_i)^2}\,dw_1 \ldots dw_Z = \frac{k+1}{Zk+1}.$$

The Poisson distribution is the special case of the negative binomial distribution in which k tends to infinity. Under this condition, $\lambda = 1/Z$. This is as we would expect, since the Poisson distribution arises from a population in which all groups are equally represented, and so the probability that two individuals chosen at random will be found to belong to the same group must be $1/Z$.

The other extreme case of the negative binomial is the logarithmic population, which is obtained by letting Z tend to infinity and k tend to zero simultaneously so that the product Zk remains finite and tends to a quantity called α. (This is not quite the same derivation as that used by Fisher[2], but the quantity α is the same as his index of diversity.) The value obtained for λ under this limiting process is $1/(\alpha + 1)$.

It will be noticed that this last value is not consistent with the equation given by Williams[3], namely, that Yule's characteristic had the value $1{,}000/\alpha$ when applied to the logarithmic distribution. His result was obtained by applying Yule's formula to a series of expected values, whereas the present procedure is equivalent to applying the formula first and then averaging the result. Some support for the new equation is found by considering the ranges of the variables concerned. Since the characteristic cannot exceed 1,000, the earlier equation would deny to α all values less than 1; but the present one allows it the range $0 \leqslant \alpha \leqslant \infty$, while $1 \geqslant \lambda \geqslant 0$.

E. H. SIMPSON
3 West End Avenue,
Pinner.
Jan. 29.

1 Yule, "Statistical Study of Literary Vocabulary" (Cambridge, 1944).
2 Fisher, Corbet and Williams, *J. Animal Ecol.*, 12, 42 (1943).
3 Williams, *Nature*, 157, 482 (1946).