MEASUREMENT OF DIVERSITY

F. N. DAVID
UNIVERSITY OF CALIFORNIA, RIVERSIDE

1. Introduction

The several measurements used by ecologists to measure diversity in plant and animal populations have been summarized by Pielou [6]. This present paper is concerned with an extension of the idea of diversity in plant populations and in particular with the description of data produced by a densitometer. Further papers applying the present ideas to actual forests counts where there is more than one observation to a cell will appear elsewhere.

2. The problem

A film is taken by an airplane flying over a natural forest. The film is put through a densitometer which prints out at equal intervals a letter corresponding to its optical density. In the particular experiment which was presented to us there were 120 letters printed out for the scan across the film, the letters being A through G inclusive. The number of letters for the scan down the film is dependent only on the length of the film. The optical density of the film and therefore the letter corresponding to it is supposedly representative of the type of tree. A measure of the clustering of the trees is required.

Essentially the same problem arises if the forest is gridded, the fuel bed computed for each square, and the results of the computations assigned to one or other of ten classes.

It will be recognized that if there are m letters one way and n letters the other the problem reduces to that of a board with $m \times n$ cells on which $m \times n$ letters are arranged, one letter to a cell, the arrangement under the null hypothesis being that of randomness.

3. Notation

Let there be s kinds of letters, with k_t of the tth kind. Denote

$$\sum_{t=1}^{s} k_t^{(r)} = K_r \tag{3.1}$$

and

$$K_1 = mn = \sum_{t=1}^{s} k_t. \tag{3.2}$$

With the partial support of the National Institutes of Health

Consider the lattice of $(m - 1)(n - 1)$ lines formed by the boundaries of the cells, excluding the border framework. Define a random variable $\alpha_{i,j}$, $i = 1, \cdots, m - 1; j = 1, \cdots, n - 1$, associated with the i,jth node of the lattice.

Let $\alpha_{i,j}$ score 6 if all four letters surrounding the node are the same,
score 3 if three letters are the same and one different,
score 2 if two are the same and two are the same,
score 1 if two are the same and two different,
score 0 if all four are different.

(Other methods of scoring are possible and will be discussed later.)

We propose

$$S = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \alpha_{i,j} \tag{3.3}$$

as a measure of clustering and of diversity.

4. Properties of S

The basic hypothesis is that the $m \times n$ letters are randomly placed in the $m \times n$ squares. The alternate hypothesis, as far as forestry problems of position, of disease, and so on, are concerned is that letters of like kind tend to cluster together. Consequently, a large value of S will indicate possibly that there is clustering, a small value will indicate that like letters are more widely dispersed from each other than might be anticipated. From the point of view of the forestry problems the acceptance or rejection of an hypothesis of randomness is not important. The field ecologist is concerned chiefly in the calculation of an index of diversity, range zero to unity, which will be large when the numbers of species are equal and the species are clumped together, and small when the converse is true. The maximum and minimum values of S are reasonably easy to compute so that one possible index is

$$I_1 = \frac{S - \min S}{\max S - \min S}. \tag{4.1}$$

From a statistical point of view this is not very satisfactory since the distribution of S may have very long tails in which case I_1 could give a very misleading result. We have tended up to the present to use

$$I_2 = \frac{S - E(S)}{\sigma_S} \tag{4.2}$$

which, although familiar to the statistician, is not liked by the biologist because it has not got a zero-one range. Undoubtedly in the present writer's opinion the best index would be to compute

$$I_3 = P\{S \leqq S_o\}, \tag{4.3}$$

where S_o is the observed value. But to compute I_3 the distribution of S must be known and this for the present we do not have. The algebraic derivation of the mean and variance of S is given in succeeding sections.

5. Algebraic attack

For the mean and variance of S under any system of scoring we rely heavily on the tables of Augmented Symmetric Functions constructed by David and Kendall [3]. The algebra is elementary and simple in principle. For example, it is clear that we may write

$$K_1^{(4)} = \sum_{\ell=1}^{s} k_\ell^{(4)} + 4\sum_{\ell\neq h} k_\ell^{(3)}k_h + 3\sum_{\ell\neq h} k_\ell^{(2)}k_h^{(2)} + 6\sum_{\ell\neq h\neq t} k_\ell^{(2)}k_hk_t + \sum_{\ell\neq h\neq t\neq r} k_\ell k_h k_t k_r, \tag{5.1}$$

formally corresponding to the expansion of power sums in terms of the Augmented Monomial Symmetric Functions (David and Kendall [3]), namely,

$$(1)^4 = \sum_{\ell=1}^{s} k_\ell^4 + 4\sum_{\ell\neq h} k_\ell^3k_h + 3\sum_{\ell\neq h} k_\ell^2k_h^2 + 6\sum_{\ell\neq h\neq t} k_\ell^2k_hk_t + \sum_{\ell\neq h\neq t\neq r} k_\ell k_h k_t k_r = [4] + 4[31] + 3[2^2] + 6[21^2] + [1^4]. \tag{5.2}$$

This raises the possibility of expressing what may be called the Augmented Factorial Monomial Symmetric Functions (AFMSF for short) in terms of powers and products of the K and vice versa, with numerical coefficients which are the same as those in the power sum AMSF relationship. The product of the K-functions is easy to compute algebraically but is difficult to express in any succinct fashion. Thus if we write

$$\sum_{\ell\neq h\neq t} k_\ell^{(a)}k_h^{(b)}k_t^{(c)} = [k_ak_bk_c], \tag{5.3}$$

the first relationship given above is, symbolically,

$$K_1^{(4)} = [k_4] + 4[k_3k_1] + 3[k_2^2] + 6[k_2k_1^2] + [k_1^4], \tag{5.4}$$

but few of the others show such simplicity. Table I gives the formal correspondence between some of the power sums and the K-functions, w being the

TABLE I

K-FUNCTIONS CORRESPONDING TO GIVEN POWER SUMS

Power Sum	K-Functions
$(n)(1)^{w-n}$	$K_n(K_1 - n)^{(w-n)}$
$(n)(2)$	$K_n(K_2 - n^{(2)}) - 2nK_{n+1}$
$(n)(2)^2$	$K_n(K_2 - n^{(2)})(K_2 - n^{(2)} - 2) - 4nK_{n+1}(K_2 - n^{(2)} - 2) - 4K_n(K_3 - n^{(3)}) + 4n(n+3)K_{n+2} + 16n^{(2)}K_{n+1}$
$(2)^2(1)^{w-4}$	$(K_2(K_2 - 2) - 4K_3)(K_1 - 4)^{(w-4)}$
$(3)^2(1)^{w-6}$	$(K_3(K_3 - 6) - 9K_5 - 18K_4)(K_1 - 6)^{(w-6)}$
$(4)^2(1)^{w-8}$	$(K_4(K_4 - 24) - 16K_7 - 72K_6 - 96K_5)(K_1 - 8)^{(w-8)}$

weight. Generally we will have, corresponding to $(m)(n)$, the K-function

$$K_n(K_m - n^{(m)}) - {}^mC_1{}^nC_1K_{n+m-1} - 2!{}^mC_c{}^nC_2K_{n+m-2} - \cdots - {}^mC_{m-1}{}^nC_{m-1}(m-1)!K_{n+1}. \tag{5.5}$$

These enable Tables AI of the appendix to be written down immediately from the tables already in existence. Further relationships between the K-products and the AFMSF were worked out and tables of higher weights were constructed. These however—and the complete tables—proved to be unnecessary for our particular problem and so we do not reproduce them here.

6. Mean of S (one, two, and three dimensions)

It is clear that the two dimensional problem described earlier is just a special case of scoring over a lattice in any number of dimensions. The method used, involving the use of the AFMSF is applicable for any number of dimensions and any method of scoring.

6.1. *One dimension.* For one dimension the problem is that of the multicolored run which has been completely solved (Barton and David [1]). Define

$$\alpha_i = \begin{cases} 1 & \text{if same color either side of } i\text{th gap} \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Then

$$S = \sum_{i=1}^{K_1-1} \alpha_i. \tag{6.2}$$

We have

$$K_1^{(2)} = \sum_{\ell=1}^{s} k_\ell^{(2)} + \sum_{\ell \neq v} k_\ell k_v, \tag{6.3}$$

$$E(\alpha_i) = \frac{1}{K_1^{(2)}}\left[1 \sum_{\ell=1}^{s} k_\ell^{(2)} + 0 \sum_{\ell \neq v} k_\ell k_v\right] = \frac{K_2}{K_1^{(2)}} \tag{6.4}$$

and

$$ES = \frac{K_2}{K_1}. \tag{6.5}$$

6.2. *Two dimensions.* Corresponding to the (i,j) node of an $mn(=K_1)$ board, we have

$$S = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \alpha_{i,j}. \tag{6.6}$$

Starting from the relationship

$$K_1^{(4)} = \sum_\ell k_\ell^{(4)} + 4 \sum_{\ell \neq v} k_\ell^{(3)} k_v + 3 \sum_{\neq} k_\ell^{(2)} k_v^{(2)} + 6 \sum_{\neq} k_\ell^{(2)} k_v k_h + \sum_{\neq} k_\ell k_v k_h k_r, \tag{6.7}$$

each permutation of a of one color, b of another, and so on ($a + b + \cdots = 4$) will give rise to the same score. So we may write

$$(6.8)\qquad E(\alpha_{i,j}) = \frac{1}{K_1^{(4)}}\Big[s_4 \sum_{\ell} k_\ell^{(4)} + 4s_3 \sum_{\ell \neq v} k_\ell^{(3)} k_v + 3s_2 \sum_{\neq} k_\ell^{(2)} k_v^{(2)} + 6s_1 \sum_{\neq} k_\ell^{(2)} k_v k_n + s_0 \sum_{\neq} k_\ell k_v k_n k_r\Big].$$

For the scoring system given previously (Section 3)

$$(6.9)\qquad s_4 = 6, \quad s_3 = 3, \quad s_2 = 2, \quad s_1 = 1, \quad s_0 = 0,$$

and

$$(6.10)\qquad E(\alpha_{i,j}) = \frac{6}{K_1^{(4)}}\Big[\sum_{\ell} k_\ell^{(4)} + 2 \sum_{\ell \neq v} k_\ell^{(3)} k_v + \sum_{\neq} k_\ell^{(2)} k_v^{(2)} + \sum_{\neq} k_\ell^{(2)} k_v k_h\Big].$$

It is possible to write each of the AFMSF in terms of sums of products of the K-functions, using the upper half of Table AI in the appendix, but it is easier to note that the expression in the brackets is the expansion of $K_2(K_1 - 2)^{(2)}$. Accordingly

$$(6.11)\qquad E(\alpha_{i,j}) = \frac{6K_2}{K_1^{(2)}}.$$

An alternative method of scoring is not to count diagonal elements. Thus we score

$$\begin{matrix} A\;A \\ \cdot \\ A\;A \end{matrix}\; 4, \quad \begin{matrix} A\;A \\ \cdot \\ A\;B \end{matrix}\; 2, \quad \begin{matrix} A\;B \\ \cdot \\ A\;B \end{matrix}\; 2, \quad \begin{matrix} A\;B \\ \cdot \\ B\;A \end{matrix}\; 0, \quad \begin{matrix} A\;B \\ \cdot \\ A\;C \end{matrix}\; 1,$$

$$\begin{matrix} A\;C \\ \cdot \\ B\;A \end{matrix}\; 0, \quad \begin{matrix} A\;C \\ \cdot \\ B\;D \end{matrix}\; 0.$$

This will mean that not all permutations, for given $a, b, c, \cdots$, and so on, will have the same weight and we will have

$$(6.12)\qquad E(\alpha_{i,j}) = \frac{1}{K_1^{(4)}}\Big[4 \sum k_1^{(4)} + 8 \sum_{\ell \neq v} k_\ell^{(3)} k_v + 4 \sum_{\neq} k_\ell^{(2)} k_v^{(2)} + 4 \sum_{\neq} k_\ell^{(2)} k_v k_h\Big] = \frac{4K_2}{K_1^{(2)}}.$$

These two methods of scoring are the only ones that have been worked out in some detail but clearly any other method of scoring is easily applied.

6.3. *Three dimensions.* Corresponding to the (i, j, ℓ) node of an $m \times n \times t (= K_1)$ rectangular parallelepiped built up from unit cubes we have

$$(6.13)\qquad S = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \sum_{\ell=1}^{t-1} \alpha_{i,j,\ell}.$$

The fundamental relationship is

$$K_1^{(8)} = \sum_a k_a^{(8)} + 8 \sum_{\neq} k_a^{(7)} k_b + 28 \sum_{\neq} k_a^{(6)} k_b^{(2)} + \cdots + \sum_{\neq} k_a k_b k_c k_d k_e k_f k_g k_h. \tag{6.14}$$

If diagonal joins between like letters are allowed then the score for each partition, and therefore for each permutation, is written down immediately and we have

$$E(\alpha_{i,j,\ell}) = \frac{28 K_2}{K_1^{(2)}}. \tag{6.15}$$

If diagonal joins between like letters are not allowed then the score has to be found for each permutation within a given partition and

$$E(\alpha_{i,j,\ell}) = \frac{12 K_2}{K_1^{(2)}}. \tag{6.16}$$

It proved more convenient in subsequent manipulations to denote by Z the quantity $K_2(K_1 - 2)^{(2)}$. We have then for mean S

one dimension	$Z(K_1 - 1)/K_1^{(4)}$;	
two dimensions	$6Z(m - 1)(n - 1)/K_1^{(4)}$,	diagonal joins allowed;
	$4Z(m - 1)(n - 1)/K_1^{(4)}$,	diagonal joins not allowed;
three dimensions	$28Z(m - 1)(n - 1)(t - 1)/K_1^{(4)}$,	diagonal joins allowed;
	$12Z(m - 1)(n - 1)(t - 1)/K_1^{(4)}$,	diagonal joins not allowed.

7. Free sampling

The algebraic approach for obtaining the expected value of S is conditional on the $\{k_\ell\}$, $\ell = 1, \cdots, s$. Some simplification in the formulae results if it is supposed that the k are obtained as the result of some sampling procedure either, say, free multinomial sampling or a Pólya multiple urn process. Suppose K_1 letters are drawn from an urn in which the proportion of the ℓth kind is kept constant and eoual to p_ℓ, and

$$\sum_{\ell=1}^{s} p_\ell = 1. \tag{7.1}$$

These K_1 letters are put down randomly as before. We have then, if we write

$$\sum_{\ell=1}^{s} p_\ell^r = P_r, \tag{7.2}$$

that

(7.3) one dimension $E(S) = (K_1 - 1)P_2$;

two dimensions

(7.4) $$E(S)\begin{cases} = 6(m-1)(n-1)P_2, & \text{diagonal joins allowed,} \\ = 4(m-1)(n-1)P_2, & \text{diagonal joins not allowed;} \end{cases}$$

three dimensions

(7.5) $$E(S)\begin{cases} = 28(m-1)(n-1)(t-1)P_2, & \text{diagonal joins allowed,} \\ = 12(m-1)(n-1)(t=1)P_2, & \text{diagonal joins not allowed.} \end{cases}$$

If we use a Pólya model where for each letter the proportion of letters returned to the urn after drawing is the same and is equal to δ, then, writing

(7.6) $$P_2 = \sum_{\ell=1}^{s} \frac{p_\ell(p_\ell + \delta)}{1(1+\delta)},$$

the expectations are formally the same as above.

8. Second moment of S

The second moment of S represents no intrinsic difficulty although the enumeration of the scores for the different permutations within a given partition allows scope for error.

8.1. *One dimension.* Here

(8.1) $$S^2 = \sum_{i=1}^{K_1-1} \alpha_i^2 + 2\sum_{i=1}^{K_1-2} \alpha_i\alpha_{i+1} + 2\sum_{i=1}^{K_1-3}\sum_{j=i+2}^{K_1-1} \alpha_i\alpha_j$$

and if we write $Y = K_3(K_1 - 3)$, $X = K_2(K_2 - 2) - 4K_3$, with Z as before, we have

(8.2) $$E(\alpha_i^2) = \frac{K_2}{K_1^{(2)}} = \frac{Z}{K_1^{(4)}},$$

(8.3) $$E(\alpha_i\alpha_{i+1}) = \frac{1}{K_1^{(3)}}\left[1\sum_{a=1}^{s} k_a^{(3)} + (2)(0)\sum_{a\neq b} k_a^{(2)}k_b + (1)(0)\sum_{a\neq b\neq c} k_a k_b k_c\right]$$

$$= \frac{K_3}{K_1^{(3)}} = \frac{Y}{K_1^{(4)}},$$

(8.4) $$E(\alpha_i\alpha_j) = \frac{1}{K_1^{(4)}}\left[1\sum k_a^{(4)} + (4)(0)\sum k_a^{(3)}k_b + (3)\left(\frac{1}{3}\right)\sum k_a^{(2)}k_b^{(2)} \right.$$

$$\left. + (6)(0)\sum k_a^{(2)}k_b k_c + (1)(0)\sum k_a k_b k_c k_d\right],$$

$$= \frac{K_2(K_2-2) - 4K_3}{K_1^{(4)}} = \frac{X}{K_1^{(4)}}.$$

Accordingly,

$$E(S^2) = \frac{1}{K_1^{(4)}}[(K_1 - 1)Z + 2(K_1 - 2)Y + (K_1 - 2)(K_1 - 3)X]. \tag{8.5}$$

It may be noted for purposes of symmetry with the larger dimensions that

$$(K_1 - 2)(K_1 - 3) = (K_1 - 1)^2 - (3K_1 - 5). \tag{8.6}$$

8.2. *Two dimensions.* Following previous notation

$$S^2 = \left(\sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \alpha_{i,j}\right)^2. \tag{8.7}$$

The expansion of the square of this double sum may be written out as in Table II.

TABLE II

NUMBER OF TERMS IN SUMMATIONS OF (8.7)

	i^2	$i(i+1)$	$i, h, (\lvert i-h \rvert \geqq 2)$	Totals
j^2	$(m-1)(n-1)$	$2(m-2)(n-1)$	$(m-2)^{(2)}(n-1)$	$(m-1)^2(n-1)$
$j(j+1)$	$2(m-1)(n-2)$	$4(m-2)(n-2)$	$2(m-2)^{(2)}(n-2)$	$2(m-1)^2(n-2)$
$j, r (\lvert j-r \rvert \geqq 2)$	$(m-1)(n-2)^{(2)}$	$2(m-2)(n-2)^{(2)}$	$(m-2)^{(2)}(n-2)^{(2)}$	$(m-1)^2(n-2)^{(2)}$
Totals	$(m-1)(n-1)^2$	$2(m-2)(n-1)^2$	$(m-2)^{(2)}(n-1)^2$	$(m-1)^2(n-1)^2$

There are four types of terms for which it will be necessary to calculate expectations.

(i) $\alpha_{i,j}^2$, $(m-1)(n-1)$ *terms.* From the expansion for $K_1^{(4)}$, allowing diagonal joins, we have

$$E(\alpha_{i,j}^2) = \frac{1}{K_1^{(4)}}[36 \sum k^{(4)} + 36 \sum k^{(3)}k + 12 \sum k^{(2)}k^{(2)} + 6 \sum k^{(2)}kk] \tag{8.8}$$

This may be split up as follows

$$\begin{aligned} 6K_1(K_1 - 2)^{(2)} &= 6[k_4] + 12[k_3k_1] + 6[k_2^2] + 6[k_2k_1^2] \\ 6K_2(K_2 - 2) - 4K_3 &= 6[k_4] \qquad\qquad\qquad + 6[k_2^2] \\ 24K_3(K_1 - 3) &= 24[k_4] + 24[k_3k_1] \end{aligned} \tag{8.9}$$

so that

$$E(\alpha_{i,j}^2) = \frac{1}{K_1^{(4)}}[6X + 24Y + 6Z] \tag{8.10}$$

If diagonal joins are not allowed

$$E(\alpha_{i,j}^2) = \frac{1}{K_1^{(4)}}[4Z + 8Y + 4Z]. \tag{8.11}$$

(ii) $\alpha_{i,j}\alpha_{i,j+1}$, *$2(m-1)(n-2)$ terms.* The expectation will be the same for $\alpha_{i,j}\alpha_{i+1,j}$ although the number of terms will be $2(m-2)(n-1)$. We start from the expansion of $K_1^{(6)}$ in terms of the AFMSF. For each permutation of a given partition it is necessary to calculate the score having regard to the fact that two squares or cells are held in common by each of the other two. Table AII of the appendix gives the total score for each permutation with diagonals allowed and not allowed. Using this table we have for example

$$(8.12)\qquad E(\alpha_{i,j}\alpha_{i,j+1}) = \frac{1}{K_1^{(6)}}\{36[k_6] + 90[k_5k_1] + 112[k_4k_2] + 58[k_3^2] + 73[k_4k_1^2] + 140[k_3k_2k_1] + 22[k_2^3] + 20[k_3k_1^3] + 25[k_2^2k_1^2] + [k_2k_1^4]\}.$$

This may be split up to give Table III, using the formal AMSF correspondence

TABLE III

ILLUSTRATION OF CALCULATION OF THE EXPECTATION OF $\alpha_{i,j}\alpha_{i,j+1}$

	$[k_6]$	$[k_5k_1]$	$[k_4k_2]$	$[k_4k_1^2]$	$[k_3^2]$	$[k_3k_2k_1]$	$[k_3k_1^3]$	$[k_2^3]$	$[k_2^2k_1^2]$,	$[k_2k_1^4]$
$Z(K_1-4)^{(2)}$	1	4	7	6	4	16	4	3	6	1
$19X(K_1-4)^{(2)}$	19	38	57	19	38	76	·	19	19	
$16Y(K_1-4)^{(2)}$	16	48	48	48	16	48	16			
	36	90	112	73	58	140	20	22	25	1

(Table I·1·6 [4]), so that

$$(8.13)\qquad E(\alpha_{i,j}\alpha_{i,j+1}) = \frac{1}{K_1^{(4)}}[19X + 16Y + Z] = E(\alpha_{i,j}\alpha_{i+1,j}).$$

If diagonals are not allowed then

$$(8.14)\qquad E(\alpha_{i,j}\alpha_{i,j+1}) = \frac{1}{K_1^{(4)}}[9X + 6Y + Z] = E(\alpha_{i,j}\alpha_{i+1,j}).$$

The process is essentially the same for the other two expectations required. For $E(\alpha_{i,j}\alpha_{i+1,j+1})$ we start from $K_1^{(7)}$ and calculate the score for each partition for two sets of four squares which have one square in common. For $E(\alpha_{i,j}\alpha_{\ell,h})$ we start from $K_1^{(8)}$ scoring two sets of four squares which do not have a square in common. The results are

$$(8.15)\qquad \begin{aligned} E(\alpha_{i,j}\alpha_{i+1,j+1}) &= \frac{1}{K_1^{(4)}}[27X + 9Y], &\quad \frac{1}{K_1^{(4)}}[12X + 4Y],\\ E(\alpha_{i,j}\alpha_{\ell,h}) &= \frac{1}{K_1^{(4)}}[36X], &\quad \frac{1}{K_1^{(4)}}[16X], \end{aligned}$$

the first expression being where diagonals are allowed and the second where diagonals are not allowed in each case.

While the expressions for the expectations of products of the α may be written comparatively simply, no real simplicity appears possible for the second moment (or the variance) of S. We have

if diagonals are allowed

$$(8.16)\quad E(S^2) = \frac{1}{K_1^{(4)}}\{(m-1)(n-1)[6X+24Y+6Z] + [2(m-2)(n-1)+2(m-1)(n-2)][19X+16Y+Z] + 4(m-2)(n-2)[27X+9Y] + [(m-1)^2(n-1)^2 - (3m-5)(3n-5)][36X]\},$$

and if diagonals are not allowed

$$(8.17)\quad E(S^2) = \frac{1}{K_1^{(4)}}\{(m-1)(n-1)[4X+8Y+4Z] + [2(m-2)(n-1)+2(m-1)(n-2)][9X+6Y+Z] + 4(m-2)(n-2)[12X+4Y] + [(m-1)^2(n-1)^2 - (3m-5)(3n-5)][16X]\}.$$

The simplicity which resulted, in the one dimensional case, from calculating the second factorial moment of S does not follow for two or three dimensions.

8.3. *Three dimensions.* For three dimensions we have

$$(8.18)\quad E(\alpha_{i,j,h}^2) = \frac{1}{K_1^{(4)}}[420X+336Y+28Z], \qquad \frac{1}{K_1^{(4)}}[84X+48Y+12Z],$$

$$(8.19)\quad E(\alpha_{i,j,h}\alpha_{i,j,h+1}) = \frac{1}{K_1^{(4)}}[594X+184Y+6Z], \qquad \frac{1}{K_1^{(4)}}[112X+28Y+4Z],$$

$$(8.20)\quad E(\alpha_{i,j,h}\alpha_{i,j+1,h+1}) = \frac{1}{K_1^{(4)}}[687X+96Y+Z], \qquad \frac{1}{K_1^{(4)}}[119X+24Y+Z],$$

$$(8.21)\quad E(\alpha_{i,j,h}\alpha_{i+1,j+1,h+1}) = \frac{1}{K_1^{(4)}}[735X+49Y], \qquad \frac{1}{K_1^{(4)}}[135X+9Y],$$

$$(8.22)\quad E\alpha_{i,j,h}\alpha_{a,b,c}) = \frac{1}{K_1^{(4)}}[784X], \qquad \frac{1}{K_1^{(4)}}[144X].$$

The expressions to the left are those in which diagonal counts are allowed and those to the left are those where they are not allowed. The numbers of terms are, respectively,

$$(m-1)(n-1)(t-1),$$
$$2(m-1)(n-1)(t-2)+2(m-1)(n-2)(t-1)+2(m-2)(n-1)(t-1),$$
$$4(m-1)(n-2)(t-2)+4(m-2)(n-1)(t-2)+4(m-2)(n-2)(t-1),$$
$$8(m-2)(n-2)(t-2),$$
$$(m-1)^2(n-1)^2(t-1)^2-(3m-5)(3n-5)(3t-5).$$

The second moment of S follows by simple multiplication of these expressions but, since no simplification results, I have not formally written it out.

9. Second moment under free sampling

The quantities intervening in the second moment which have to be taken into account in passing from conditional to unconditional expectations are X, Y, Z. The two latter present no difficulty:

$$E(Y)=K_1^{(4)}P_3,\qquad E(Z)=K_1^{(4)}P_2, \tag{9.1}$$

while

$$X=K_2^2-2K_2-4K_3=\sum_\ell k_{2\ell}k_{2\ell}+\sum_{\neq}k_{2\ell}k_{2h}-2\sum k_{2\ell}-4\sum k_{3\ell} \tag{9.2}$$
$$=\sum k_{4\ell}+\sum_{\neq}k_{2\ell}k_{2h}$$

and

$$E(X)=[P_4+\sum_{\neq}p_1^2p_h^2]K_1^{(4)}=P_2^2K_1^{(4)}. \tag{9.3}$$

Accordingly, the unconditional expectation of the second moment of S will be obtained by substituting P_2, P_3, and P_2^2 for $Z/K_1^{(4)}$, $Y/K_1^{(4)}$, and $X/K_1^{(4)}$, respectively. The variance of S (2D, diagonals) may be written as

$$\operatorname{Var}S=(m-1)(n-1)[10P_2+124P_3-134P_2^2] \tag{9.4}$$
$$-[(m-1)+(n-1)][2P_2+68P_3-70P_2^2]+36[P_3-P_2^2].$$

For the other cases we have

$$\operatorname{Var}S\,(\text{2D, no diagonals})=(m-1)(n-1)[8P_2+48P_3-56P_2^2] \tag{9.5}$$
$$-[(m-1)+(n-1)][2P_2+28P_3-30P_2^2]+16[P_3-P_2^2],$$

$$\operatorname{Var}S\,(\text{3D, diagonals}) \tag{9.6}$$
$$=(m-1)(n-1)(t-1)[76P_2+2984P_3-3060P_2^2]$$
$$-[(m-1)(n-1)+(n-1)(t-1)$$
$$+(t-1)(m-1)][20P_2+1528P_3-1548P_2^2]$$
$$+[(m-1)+(n-1)+(t-1)][4P_2+776P_3-780P_2^2]$$
$$-392[P_3-P_2^2],$$

(9.7) Var S (3D, no diagonals)

$$\begin{aligned} &= (m-1)(n-1)(t-1)[48P_2 + 576P_3 - 624P_2^2] \\ &\quad - [(m-1)(n-1) + (n-1)(t-1) \\ &\quad + (t-1)(m-1)][16P_2 + 320P_3 - 336P_2^2] \\ &\quad + [(m-1) + (n-1) + (t-1)][4P_2 + 168P_3 - 172P_2^2] \\ &\quad - 72[P_3 - P_2^2]. \end{aligned}$$

10. Special cases

(i) When all the probabilities are the same, that is, when

$$p_\ell = \frac{1}{s} \qquad \ell = 1, \cdots, s, \tag{10.1}$$

the last terms of each expression are zero and $(s-1)/s^2$ is a factor. Thus, in order of writing above we have

$$\frac{s-1}{s^2}\{10(m-1)(n-1) - 2[(m-1) + (n-1)]\},$$

$$\frac{s-1}{s^2}\{8(m-1)(n-1) - 2[(m-1) + (n-1]\},$$

$$\begin{aligned} \frac{s-1}{s^2}\{76(m-1)(n-1)(t-1) &- 20[(m-1)(n-1) + (m-1)(t-1) \\ &+ (n-1)(t-1)] + 4[(m-1) + (n-1) + (t-1)]\}, \end{aligned}$$

$$\begin{aligned} \frac{s-1}{s^2}\{48(m-1)(n-1)(t-1) &- 16[(m-1)(n-1) + (m-1)(t-1) \\ &+ (n-1)(t-1)] + 4[(m-1) + (n-1) + (t-1)]\}. \end{aligned}$$

(ii) When there are only two kinds of letters

$$p_1 = p, \qquad p_2 = 1 - p_1 = q, \qquad p_3 = \cdots = p_s = 0 \tag{10.2}$$

and

$$P_2 = 1 - 2pq, \qquad P_3 = 1 - 3pq. \tag{10.3}$$

In order of writing the variances now become

$$\begin{aligned} pq[(m-1)(n-1)(144 - 536pq) - [(m-1) + (n-1][72 - 280pq] \\ + 36(1 - 4pq)], \end{aligned}$$

$$\begin{aligned} pq[(m-1)(n-1)(64 - 224pq) - [(m-1 + (n-1)][32 - 120pq] \\ + 16(1 - 4pq)], \end{aligned}$$

$$\begin{aligned} &pq[(m-1)(n-1)(t-1)[3136 - 12240pq] - [(m-1)(n-1) \\ &\quad + (n-1)(t-1) + (t-1)(m-1)][1568 - 6192pq] \\ &\quad + [(m-1) + (n-1) + (t-1)][748 - 3120pq] - 392[1 - 4pq], \end{aligned}$$

$$pq[(m-1)(n-1)(t-1)[672-2469pq]-[(m-1)(n-1) + (n-1)(t-1)+(t-1)(m-1)][352-1344pq] + [(m-1)+(n-1)+(t-1)][176-688pq]-72(1-4pq)].$$

11. Second moment under a multiple Pólya model

Let

$$(11.1)\qquad P\{k_i\} = \frac{K_1!}{\prod_{i=1}^{s} k_i!} \frac{\prod_{j=1}^{s}\prod_{i=0}^{k_j-1}(p_j+i\delta)}{\prod_{h=0}^{K_1-1}(1+h\delta)}$$

($\delta = 0$ gives the multinomial). Taking for example the two dimensional case, diagonals allowed, we have

$$(11.2)\qquad E(S) = \frac{6(P_2+\delta)(m-1)(n-1)}{1+\delta}$$

and writing

$$(11.3)\qquad D_4 = 1(1+\delta)(1+2\delta)(1+3\delta),$$

we have for the unconditional variance of S

$$(11.4)\ \operatorname{Var} S = \frac{1}{D_4(1+\delta)}[(m-1)^2(n-1)^2\{144\,\delta(P_3-P_2^2) + 72\,\delta^2(P_2+2P_3-3P_2^2)+72\,\delta^3[1-P_2]\} + (m-1)(n-1)\{(10P_2+124P_3-134P_2^2) + \delta(10+164P_2-40P_3-134P_2^2)+\delta^2(174-10P_2-164P_3) + \delta^3(164-164P_2)\} + ((m-1)+(n-1))\{(2P_2+68P_3-70P_2^2) + \delta(2+76P_2-8P_3-70P_2^2)+\delta^2(72-2P_2-70P_3) + \delta^3(76-76P_2)\} + 36\{(P_3-P_2^2)-\delta(P_2-P_3)+\delta^2(1-P_3)+\delta^3(1-P_2)\}].$$

The results for other methods of scoring and for other dimensions follow similarly.

12. Higher moments of S

The higher moments of S present no intrinsic difficulty although a certain amount of counting is required. It may be noted that the third moment will involve X, Y, Z, and

$$(12.1)\qquad \begin{aligned} W &= K_2(K_2-2)(K_2-4)-12K_3(K_2-2)+40(K_4+K_3),\\ V &= K_3(K_2-6)-6K_4, \qquad\qquad U = K_4,\end{aligned}$$

but no others. The fourth moment will involve the previous six quantities and

$$(12.2)\qquad \begin{aligned} R = {} & K_2(K_2 - 2)(K_2 - 4)(K_2 - 6) - 24K_3(K_2 - 3)(K_2 - 5) \\ & + 112K_3(K_2 - 3) + 160K_4(K_2 - 4) + 48K_3(K_3 - 3) - 672K_5 \\ & - 1616K_4 - 408K_3, \end{aligned}$$

$$T = K_3(K_2 - 3)(K_2 - 5),\ Q = K_4(K_2 - 4),\ M = K_3(K_3 - 3),\ N = K_5.$$

13. Examples

A knowledge of the movements of S will enable an approximation to its distribution to be made and therefore a calculation of the index of diversity I suggested earlier. The distribution of S will depend on the size and configuration of the lattice and the letter specification, as will be illustrated below. The difficulty is to find a functional form sufficiently flexible to take account of all these factors, but two approaches are possible.

EXAMPLE 1. Consider a letter specification (12,4) on a 4×4 tableau, with the diagonals counting in the scoring. The distribution of S is shown in Table IV.

TABLE IV

DISTRIBUTION OF S FOR A 4×4 TABLEAU WITH OBJECT SPECIFICATION (12, 4)

S	24	25	26	27	28	29	30	31	32	33	
Frequency	28	20	83	96	140	124	132	216	140	108	
S	34	35	36	37	38	39	40	41	42	43	Total
Frequency	148	180	40	108	128	40	36	32	17	4	1820

The momental constants are

$$\mu_1' = 32.4, \quad \mu_3 = 13.90735, \quad \beta_1 = 0.03811, \quad \sqrt{\beta_1} = 0.1952,$$
$$\mu_2 = 17.18505, \quad \mu_4 = 687.15027, \quad \beta_2 = 2.32675, \quad \sigma = 4.1455.$$

(i) Pearson and Hartley [5], p. 206, give the standardized deviates for Pearson curves for given percentage points. Using the (β_1, β_2) of the distribution above, we get Table V. These figures indicate that the *Biometrika* table will be good enough for significance levels if they are ever required.

TABLE V

PERCENTAGE POINTS OF S FROM PEARSON CURVES

Percentage Point	Deviate from Tables	Actual Percentages
5	39.546	4.9
2.5	40.583	2.9
1	41.649	1.15

(ii) Fisher-Cornish [2] give an inverse Edgeworth expansion for the percentage points of a distribution with known momental constants. Using their expansion we have Table VI. The deviates obtained by the Fisher-Cornish expansion do not differ markedly from those obtained from the Pearson system.

TABLE VI

PERCENTAGE POINTS OF S FROM THE EDGEWORTH SERIES

Percentage Point	Deviate	Actual Percentage
5	39.502	4.9
2.5	40.694	2.9
1	41.927	1.15
50	32.2651	46.2

EXAMPLE 2. Table AIII of the Appendix gives the distribution of S for eight letters (different specifications) on a 4×2 board. All these distributions will be, to a certain extent, atypical in that the edge effects will play an undue part. The difference between the Pearson system and the Fisher-Cornish expansion is a little more marked than for Example 1. If we take, for instance, the distribution with object specification (421^2) the (β_1, β_2) point is (0.755, 3.321) and the 5 percent points are 7.306 and 7.198 for Pearson and Fisher-Cornish, respectively. The actual percentage in each case is 4.29.

It is clear that if we obtain formulae for the third and fourth moments of S, we may approximate to its distribution reasonably well as far as the biological problem is concerned. Further, sampling experiments seem to indicate that with a large lattice the distribution of S can be approximated by a constant times χ^2. The question of the configuration of the lattice is one which is of small importance for a large lattice, but may be of importance when the numbers are small. Table AIV and Figure 1 of the Appendix illustrates this point.

14. Conclusion

It is clear from Section 13 that with a knowledge of the third and fourth moments an approximation to the distribution of S is possible although it would be idle to suggest that either of the methods suggested will be that ultimately chosen. The algebraic derivation of the third and fourth moments, as indicated earlier, is simple in principle but tedious in procedure. Some headway has been made with the third moment. If, as appears likely, it is close to the third moment of a constant times χ^2, then little would be gained from the derivation of the fourth moment.

The practical applications of the foregoing theory and the extension of this theory to the case where there is more than one letter per square, are numerous in the ecological field. It is proposed to publish these elsewhere.

The problem (and others allied to it) set out at the beginning of this paper arose from consultations at the Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture.

APPENDIX

TABLE AIa

AUGMENTED FACTORIAL MONOMIAL SYMMETRIC FUNCTIONS
WEIGHT 2

$(w = 2)$	$[k_2]$	$[k_1^2]$
K_2	**1**	-1
$K_1^{(2)}$	1	**1**

TABLE AIb

AUGMENTED FACTORIAL MONOMIAL SYMMETRIC FUNCTIONS
WEIGHT 3

$(w = 3)$	$[k_3]$	$[k_2k_1]$	$[k_1^3]$
K_3	**1**	-1	2
$K_2(K_1 - 2)$	1	**1**	-3
$K_1^{(3)}$	1	3	**1**

TABLE AIc

AUGMENTED FACTORIAL MONOMIAL SYMMETRIC FUNCTIONS
WEIGHT 4

$(w = 4)$	$[k_4]$	$[k_3k_1]$	$[k_2^2]$	$[k_2k_1^2]$	$[k_1^4]$
K_4	**1**	-1	-1	2	-6
$K_3(K_1 - 3)$	1	**1**	·	-2	8
$K_2(K_2 - 2) - 4K_3$	1	·	**1**	-1	3
$K_2(K_1 - 2)^{(2)}$	1	2	1	**1**	-6
$K_1^{(4)}$	1	4	3	6	**1**

TABLE AId

AUGMENTED FACTORIAL MONOMIAL SYMMETRIC FUNCTIONS
WEIGHT 5

$(w = 5)$	$[k_5]$	$[k_4k_1]$	$[k_3k_2]$	$[k_3k_1^2]$	$[k_2^2k_1]$	$[k_2k_1^3]$	$[k_1^5]$
K_5	**1**	-1	-1	2	2	-6	24
$K_4(K_1 - 4)$	1	**1**	·	-2	-1	6	-30
$K_3(K_2 - 6) - 6K_4$	1	·	**1**	-1	-2	5	-20
$K_3(K_1 - 3)^{(2)}$	1	2	1	**1**	·	-3	20
$[K_2(K_2 - 2) - 4K_3][K_1 - 4]$	1	1	2	·	**1**	-3	15
$K_2(K_1 - 2)^{(3)}$	1	3	4	3	3	**1**	-10
$K_1^{(5)}$	1	5	10	10	15	10	**1**

To express factorial power sums in terms of the augmented factorial symmetric functions read from left to right up to and including the heavy type diagonal. To express the augmented factorial symmetric functions in terms of the factorial power sums read vertically downwards as far as and including the heavy type diagonal.

NOTATION. For example

$$[k_3k_2] = \sum_{\neq} k_\ell^{(3)} \cdot k_h^{(2)}, \qquad K_r = \sum_{i=1}^{s} k_i^{(r)}.$$

(A.1)

$$[k_2^2k_1] = \sum_{\neq\,\neq} k_\ell^{(2)} k_h^{(2)} k_t,$$

TABLE AII

EIGHT SQUARES: TWO SQUARES IN COMMON

Number of Permutations	Partition	Total score for partition Diagonals (Yes)	Diagonals (No)
1	(6)	36	16
6	(51)	90	40
15	(42)	112	52
10	(3^2)	58	28
15	(41^2)	73	33
60	(321)	140	70
15	(2^3)	22	12
20	(31^3)	20	10
45	(2^21^2)	25	15
15	(21^4)	1	1
1	(1^6)	0	0

TABLE AIII

EXACT DISTRIBUTION OF S
$K_1 = 8, \quad m = 4, \quad n = 2$

S	Object Specification (8)	(71)	(62)	(61^2)	(53)	(521)	(51^3)	(4^2)	(42^2)	(421^2)	(41^4)	(21^6)
18	1											
17												
16												
15		4										
14			2					1				
13				2								
12		4	4	4	4							
11			8		8	4						
10			2	8		20	4	1	3			
9			8	8		4	4	8	4	4		
8			4	2	24	28		8	2	14	1	
7				4	20	44	24	8	32	24	2	
6						36		9	19	48	8	
5						32	20		96	90	8	
4							4		18	132	16	
3									36	92	26	
2										16	9	2
1												14
0												12
Totals	1	8	28	28	56	168	56	35	210	420	70	28

TABLE AIV

DISTRIBUTION OF S ACCORDING TO LATTICE CONFIGURATION IN FIGURE 1 OBJECT SPECIFICATION (42^2)

S	A	B
10	3	1
9	4	2
8	2	10
7	32	28
6	19	32
5	96	64
4	18	45
3	36	28
Total	210	210
μ'_1	5.143	5.167
μ_2	2.199	2.106
β_1	0.443	0.201
β_2	3.811	2.851

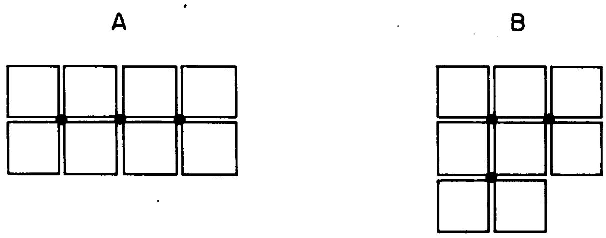

FIGURE AI

Two versions of small lattice configuration

REFERENCES

[1] D. E. BARTON and F. N. DAVID, "Multiple runs," *Biometrika*, Vol. 44 (1957), pp. 168–178.
[2] E. A. CORNISH and R. A. FISHER, "Moments and cumulants in the specifications of distributions," *Rev. Inst. Internat. Statist.*, Vol. 4 (1937) pp. 307–320.
[3] F. N. DAVID and M. G. KENDALL, "Tables of symmetric functions—part I," *Biometrika*, Vol. 36 (1949), pp. 431–449.
[4] F. N. DAVID, M. G. KENDALL, and D. E. BARTON, *Symmetric Function and Allied Tables*, Cambridge, Cambridge University Press, 1966.
[5] E. S. PEARSON and H. O. HARTLEY, *Biometrika Tables for Statisticians*, London, Cambridge University Press, 1958.
[6] E. C. PIELOU, *An Introduction to Mathematical Ecology*, New York, Wiley, 1970.