DIVERSITY AS A CONCEPT AND ITS IMPLICATIONS FOR RANDOM COMMUNITIES

G. P. PATIL and C. TAILLIE

The Pennsylvania State University, U.S.A.

1. Introduction

The results discussed in this paper are in continuation of our recent research on diversity as presented in Patil and Taillie (1976, 1977). The concept of diversity has been important in various ecological and non-ecological fields. Despite extensive literature, formal definition and logical development of diversity as a concept and its measurement have been largely lacking. Basic questions of importance are: What is diversity? Can it be measured? If so, how? and why? Our continuing research is an effort to come to grips with some of these issues.

In earlier work, we have discussed the topics of diversity ordering, sensitivity, decomposition, equitability, and estimation. We put forth the view that diversity is an average property of a community and identified that property as the species rarity. Introducing the concepts of dichotomy and ranking in this context, corresponding measures of species rarity were developed. The emphasis was on a fixed community with given abundance vector \( \pi = (\pi_1, \pi_2, \ldots) \) where \( \Sigma \pi_i = 1 \).

The present paper has two sections. The first summarizes some pertinent aspects of this earlier work and reports a new unifying formulation to measure species rarity. Various indices, including the Hurlbert-Smith index, are then obtained as special cases. The second section takes up the subject of random communities. Often it will be realistic to regard the abundance vector as a realization of some stochastic generating mechanism. In this case, our interest will focus on the "diversity producing capacity" of the mechanism and not on the diversity of particular realizations.

Two separate problems can be distinguished. First of all, we may wish only to rank random communities: to be able to say that one random community is more diverse than another. This is the ordinal problem. The problem of quantification, on the other hand, calls for a numerical measure of the diversity of random communities. Two possibilities immediately suggest

2–63
themselves: \( E[\Delta(\pi)] \) and \( \Delta(E[\pi^*]) \) where \( E \) stands for expectation and \( \Delta \) is a diversity index. These will be called the \( E\Delta \) and \( \Delta E \) measures corresponding to the diversity index \( \Delta \). However, if the abundance vector is a member of some parametric family, it may be more natural and, from the inferential standpoint, more appropriate to adopt some parameter of the model as the diversity measure. Usually this parameter will pertain specifically to the underlying model and will not fall into the framework considered in Section 1. Of course, some justification must be given for the parameter that is chosen. A central theme of Section 2 is the justification of Fisher’s “alpha” as a diversity measure associated with the (symmetric) Dirichlet model.

In addition to its attractive theoretical properties, the Dirichlet often provides an acceptable fit to empirical data. One usually finds the index \( k \) to be small and the number of species \( s \) to be large, which leads one to idealize the model by letting \( k \to 0 \) and \( s \to \infty \) with the product \( \alpha = sk \) held fixed. While the negative binomial sampling distribution converges under this limiting scheme, the symmetric Dirichlet does not. This would seem to lead to the anomaly of a sampling distribution without a valid underlying population model. The limit can be made rigorous at the population level by introducing a new concept called size-biased permutation. It is shown that the size-biased permutation of the Dirichlet does have a valid limit. Moreover, this limiting Dirichlet is itself invariant under size-biased permutation and is described in terms of iid random variables by means of what we call a residual allocation model. Further, the limiting Dirichlet is the only such residual allocation model which is invariant under size-biased permutation.

The limiting Dirichlet has but one parameter—Fisher’s \( \alpha \). We show that the result of randomly deleting a fraction \( 1-p \) of the species is still a limiting Dirichlet but with parameter \( p\alpha \). This provides some justification for interpreting \( \alpha \) as a species richness parameter.

Treating a community as random introduces an additional component of variability into the sampling distribution. In an effort to eliminate this component, there has been some interest in identifying fixed models that closely resemble realizations of the limiting Dirichlet. Engen (1975) has suggested the geometric series as a fixed version of the limiting Dirichlet but we find that this exaggerates community diversity. It is not generally recognized that the geometric series has two stochastic analogues, called preemption and dominance. The preemption mechanism generates the limiting Dirichlet but it is the dominance mechanism that has the geometric series as fixed version.
We extend the intrinsic diversity ordering to random communities. Four possible definitions are considered and these are found to be equivalent for fixed, but not for random, communities. The stochastic diversity ordering is only partial and two communities need not be comparable. However, it is shown that the pair of parameters \( (s, \alpha) \) is approximately a complete diversity measure for the Dirichlet model. Further, any two limiting Dirichlets are comparable and here \( \alpha \) is a complete diversity parameter.

Throughout the article, the terms "increasing," "decreasing," "more," and "less" have their loose interpretations unless qualified by "strictly."

2. Fixed communities

Diversity as Average Community Rarity

As discussed in Patil and Taillie (1977), we view diversity as an average property of a community. But the average of what? To an outside observer, variety is a most striking feature of a diverse community. In a diverse community, the typical species is relatively rare. Consequently, we propose that diversity be defined as the average rarity within a community. To make this idea precise, the concepts of "community" and "rarity" must be formalized.

Consider some quantity distributed among a countable set of categories, labelled \( i = 1, 2, 3, \ldots \), with \( \pi_i \) as the proportionate share received by category \( i \). This quantity may be continuous (income, biomass, etc.) or it may be discrete (word occurrences in a text, biological organisms, etc.). For concreteness, we will usually speak of biological organisms grouped into species and call \( \pi = (\pi_1, \pi_2, \pi_3, \ldots) \) the species abundance vector; arranging the components of \( \pi \) in descending order gives the ranked abundance vector, \( \pi^* = (\pi^*_1, \pi^*_2, \pi^*_3, \ldots) \) where \( \pi^*_1 > \pi^*_2 > \pi^*_3 > \ldots \). For our purposes, a community may be identified with the pair \( C = (s, \pi) \) where \( s \) is the number of non-zero components of \( \pi \), i.e., the number of species that are physically present in the community.

Given the community, a numerical measure of rarity is to be associated with each species. Denoting this rarity of species \( i \) as \( R(i; \pi) \), the diversity measure of community \( C \) is its average rarity and is given by \( \Delta(C) = \sum \pi_i R(i; \pi) \) where \( \Delta \) is the associated diversity index. Sometimes we write \( \Delta(\pi) \) instead of \( \Delta(C) \). The rarity measures considered here are of the dichotomous type: \( R(i; \pi) \) depends only on \( \pi_i \) and is denoted \( R(\pi_i) \). Without loss of generality, we have adopted the following standardization:

\[
R(1) = 0; \quad R'(1) = -1. \quad \cdots \quad (1.1)
\]

In earlier work, we have discussed a parametric family of indices \( \Delta_{\beta} = (1-\sum \pi_i^{\beta+1})/\beta \) with rarity measure \( R(\pi) = (1-\pi^\beta)/\beta \) of which the Species
Count, Shannon, and Simpson indices are special cases. The family $\Delta_\beta$ comes under discussion in this paper also.

**Hurlbert-Smith Index**

Let $\omega$ be a nonnegative integer. The Hurlbert-Smith index of order $\omega$ is the expected number of species obtained when $\omega + 1$ individuals are randomly selected from the community. The definition was given by Hurlbert (1971) and by Smith and Grassle (1977) for sampling without and with replacement, respectively. Only sampling with replacement will be considered here. It is also convenient to subtract one from the expected number of species so that a single-species community has diversity zero. With this change, the Hurlbert-Smith index takes the form

$$
\Delta^{H-S}_\omega = \sum (1 - \pi_i)(1 - (1 - \pi_i)^{\omega})
$$

$$
= s - 1 - \sum (1 - \pi_i)^{\omega+1},
$$

where the second summation ranges over the nonzero components of $\pi$. Some special cases are tabulated below:

<table>
<thead>
<tr>
<th>$\omega$</th>
<th>0</th>
<th>1</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta^{H-S}_\omega$</td>
<td>0</td>
<td>Simpson</td>
<td>$s - 1$.</td>
</tr>
</tbody>
</table>

For a given community $C$, the plot of $\Delta^{H-S}_\omega(C)$ versus $\omega$ is the expected species-individual curve. Empirical species-individual curves are generally concave and one may expect the same for the Hurlbert-Smith index.

**Theorem 1.1**: Both $\Delta_\beta(C)$ and $\Delta^{H-S}_\omega(C)$ are identically zero when $C$ is a single-species community. Otherwise, $\Delta_\beta(C)$ is a strictly decreasing, strictly convex function of $\beta$ with $\lim \Delta_\beta(C) = 0$ when $\beta \to \infty$; and $\Delta^{H-S}_\omega(C)$ is a strictly increasing, strictly concave function of $\omega$.

The shape of the empirical species-individual curve depends on the sampling scheme as well as the underlying community structure. The preceding theorem justifies a concave shape only for multinomial sampling. A similar result shows that the species-area curve is strictly increasing and strictly concave for Poisson sampling. It is perhaps well to point out that the sampling considered in the definition of the Hurlbert-Smith index is conceptual. There is no requirement that it be physically possible to draw a random sample.

The Hurlbert-Smith index is of the dichotomous type with rarity measure $R(\pi) = (1 - \pi)[1 - (1 - \pi)^{\omega}]$. In this form $R(\pi)$ seems to have no obvious interpretation. However, there is a general unifying scheme for the cons-
construction of rarity measures which yields those of both $\Delta_{\beta}$ and $\Delta_{\alpha_{-\beta}}$ after imposing the standardizing requirements (1.1).

In previous articles (Patil and Taillie (1976, 1977)), we have given an interpretation of $\Delta_{\beta}$ by considering a traveller who wanders at random through a community. Consider again such a traveller who initially encounters a member of species $i$ and subsequently encounters $X$ additional individuals where $X$ is a positive integer-valued random variable. Define the type I rarity measure to be the probability that a new species is encountered, i.e., the probability that at least one of the $X$ additional individuals belongs to a species different from $i$. A type II rarity measure, on the other hand, is the probability that each of the additional individuals belongs to species different from $i$. Clearly these probabilities are large when the species $i$ is rare.

Both types of rarity measures may be expressed in terms of the probability generating function $G(t)$ of $X$ as in Table 1(a). In their raw forms, the type I and type II rarity measures are duals of one another as defined in Patil and Taillie (1976). The type II rarity measure cannot be standardized unless $P(X = 1) > 0$, but it will be seen later that this is a necessary requirement if the associated index is to possess certain desirable properties.

**TABLE 1. RARITY MEASURES INDUCED BY RANDOM VARIABLE $X$**

(a) General

<table>
<thead>
<tr>
<th>Type</th>
<th>Raw Form</th>
<th>Standardized Form (1.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$1 - G(\pi)$</td>
<td>$[1 - G(\pi)]/G'(1) = [1 - G(\pi)]/E[X]$</td>
</tr>
<tr>
<td>II</td>
<td>$G(1 - \pi)$</td>
<td>$G(1 - \pi)/G'(0) = G(1 - \pi)/P(X = 1)$</td>
</tr>
</tbody>
</table>

(b) Degenerate

<table>
<thead>
<tr>
<th>Type</th>
<th>Raw Form</th>
<th>Standardized Form (1.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$1 - \pi^\beta$</td>
<td>$(1 - \pi^\beta)/\beta$</td>
</tr>
<tr>
<td>II</td>
<td>$(1 - \pi)^\beta$</td>
<td>Not possible unless $\beta = 1$</td>
</tr>
</tbody>
</table>

(c) Uniform Discrete

<table>
<thead>
<tr>
<th>Type</th>
<th>Raw Form</th>
<th>Standardized Form (1.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$1 - \omega^{-1} n(1 - \pi^\omega)/(1 - \pi)$</td>
<td>$2^{-1} \omega(\omega + 1) [1 - \omega^{-1} n(1 - \pi^\omega)/(1 - \pi)]$.</td>
</tr>
<tr>
<td>II</td>
<td>$\omega^{-1}(1 - \pi)[1 - (1 - \pi)^\omega]/\pi$</td>
<td>$(1 - \pi)(1 - (1 - \pi)^\omega)/\pi$</td>
</tr>
</tbody>
</table>
Particular rarity measures may now be constructed by specializing on the distribution of $X$. Three cases will be considered by first letting the distribution be degenerate, then uniform discrete and, finally, log series.

When $X$ is degenerate at the positive integer $\beta$, the generating function is $G(t) = t^\beta$ and we obtain the rarity measures of Table I(b). Here $\Delta_\beta$ is the index associated with the standardized type I measure.

Next, let $X$ be a discrete uniform random variable with support $\{1, 2, 3, \ldots, \omega\}$ and generating function $\omega^{-1}(1-t^\omega)/(1-t)$. The rarity measures appear in Table I(c). In this case, the Hurlbert-Smith index is associated with the standardized type II measure. Also, letting $\omega \to 0$ in the two raw forms gives the unfamiliar indices of Patil and Taillie (1976). Conceptually, $\omega$ is an integer but real values make mathematical sense.

Finally, let $X$ have a log series distribution with parameter $\theta$ and generating function $G(t) = \log(1-\theta t)/\log(1-\theta)$. The standardized type II rarity measure is then $R(\theta) = -\log(1+\theta\pi)/\theta$ which converges to the rarity measure of the Shannon index when $\theta \to 1$. Note that the number $X$ of additional encounters converges to infinity under this limit.

Other types of rarity measures could be defined as the probability of particular events associated with the random variable $X$. Usually such a probability turns out to be a function of the entire vector $\pi$ and not merely $\pi_i$ where $i$ is the initially encountered species. The resulting rarity measure then has the general form $R(i; \pi)$ and is not of the dichotomous type.

**Evenness and Richness**

Diversity can be loosely described as a feature of a biological community incorporating both the number of species and the evenness of their abundance. We have formalized these two aspects of diversity into a pair of operations called introducing a species and transferring abundance (see Patil and Taillie (1976) for details). The operation of introducing a species is formally a special case of transferring abundance, but the two operations are conceptually quite different. Introducing a species increases the species richness component of diversity while transferring abundance increases the evenness component.

We have also given two criteria for diversity indices. The first—called Criterion C1—applies only to dichotomous indices and is quite intuitive. It requires that the rarity measure $R(\pi)$ be a decreasing function of $\pi$ for $0 < \pi \ll 1$. Introducing a species increases the value of the index when Criterion C1 is satisfied but the same is not true for transferring abundance. Criterion
C2 requires that the index be permutation invariant and that its value increase under both operations.

Another criterion—call it C3—has been given by Lewontin (1973): the index should be a permutation invariant and concave function of the abundance vector \( \pi \). Although Criterion C3 always implies Criterion C2, it does not appear possible to directly relate concavity to the evenness and richness aspects of diversity. Criterion C3 has the disadvantage of not being preserved by increasing transformations of the index. A major advantage is the availability of Jensen’s inequality. For example, we might mention the following interesting result which is not generally true for Criterion C2.

**Theorem 1.2:** If \( \Delta(\pi) \) satisfies Criterion C3, then replacing population proportions with sample proportions results in a negatively (possibly zero) biased estimator.

Many of the diversity indices in common use satisfy Criterion C3. These include \( \Delta_\alpha \) for \( \beta \geq -1 \), \( H^{a-3}_\omega \) for \( \omega > 0 \), and all indices with a type I rarity measure. For type II rarity measures \( P(X = 1) > 0 \) is necessary for Criterion C3, while \( P(X = 1) > P(X = 2) > P(X = 3) > \ldots \) is sufficient.

**Intrinsic Diversity Ordering**

The ultimate aim of an index is the comparison of communities with respect to their diversity. However, different indices may inconsistently rank a given pair of communities, thus raising the question: when is one community more diverse than another? We have attempted to answer this question without reference to indices by defining community \( C' \) to be intrinsically more diverse than community \( C \) when \( C \) leads to \( C' \) by a finite sequence of the following operations: (i) introducing a species, (ii) transferring abundance, and (iii) relabelling species, i.e., permuting components of the abundance vector.

The ordering is only partial in that two communities may not be comparable. The choice of an index smooths the partial order into a linear order and requires a decision regarding the ranking of communities which are not intrinsically comparable. To assist in this choice, a theory of index sensitivity has been developed and related to the question of index inconsistency (of Patil and Taillie (1977)).

Solomon (1975) has proposed a diversity ordering based on the notion of majorization. Fundamental results of Hardy, Littlewood and Polya (1934) and Rado (1952) show that Solomon’s definition is equivalent to ours.
Theorem 1.3: Let \( C = (s, \pi) \) and \( C' = (s', \nu) \) be two communities. Then the following statements are equivalent:

a) \( C' \) is intrinsically more diverse than \( C \).

b) \( \Delta(C') \geq \Delta(C) \) whenever \( \Delta \) satisfies Criterion C2.

c) \( \Delta(C') \geq \Delta(C) \) whenever \( \Delta \) satisfies Criterion C3.

d) \( \pi^* \) majorizes \( \nu^* \), i.e.,

\[
\sum_{i \leq k} \pi^*_i \geq \sum_{i \leq k} \nu^*_i, \quad k = 1, 2, 3, \ldots
\]

e) \( \nu \) is a convex linear combination of permutations of \( \pi \).

The ordering, as we have defined it, is more intuitive in that it relates directly to the two aspects of diversity. Majorization, on the other hand, is easier to verify in practice. Fact e) permits the intrinsic diversity ordering to be represented on the abundance simplex (cf. Patil and Taillie (1977)).

3. Random communities

By a random community we mean simply that the abundance vector \( \pi \) is random. In general, the number of species may vary with the realization. Sometimes it is convenient to obtain \( \pi \) by normalizing a set of absolute abundances to unity.

Example 2.1: (Fixed community.) Any fixed community is also a random community whose abundance vector has a degenerate distribution.

Example 2.2: Let \( \pi \) take the values \((1, 0, 0), (.5, .5, 0), \) and \((.4, .3, .3)\), each with probability \(1/3\). In this example, the number of species varies with the realization.

Example 2.3: Let \( \pi \) assume the two values \((.5, .3, .2)\) and \((.6, .31, .09)\), each with probability \(1/2\). As examples of \( E\Delta \) and \( \Delta E \) measures, take \( \Delta \) to be Simpson’s index. One finds that \( E[\Delta(\pi)] = .5779 \) and \( \Delta(E[\pi^*]) = .58345 \).

Example 2.4: Let \( \nu \) assume the two values \((.5, .4, .1)\) and \((.2, .2, .6)\), each with probability \(1/2\). Then \( E[\nu] = (.35, .3, .35) \) and \( E[\nu^*] = (.55, .3, .15) \). Notice that \( E[\nu] \) is intrinsically more diverse than any realization of \( \nu \) and is not at all representative of \( \nu \). On the other hand, \( E[\nu^*] \) gives a better overall description of the community structure without regard to species identity. Here, \( E[\Delta_1(\nu)] = .570 \) and \( \Delta_1(E[\nu^*]) = .585 \). Note that the \( E\Delta_1 \) and \( \Delta_1E \) measures inconsistently rank \( \nu \) in comparison with the community \( \pi \) of the previous example. In each case, though, the \( E\Delta_1 \) measure is less than the \( \Delta_1E \) measure. It is a simple consequence of Jensen’s inequality that this will always be so when the index satisfies Criterion C3.
Sometimes one would like to replace a random community $\nu$ with some “typical” fixed version. The above considerations suggest that $E[\nu]$ is a poor interpretation of “typical” and that $E[\nu^\ast]$, while a somewhat better interpretation, still exaggerates community diversity. We will return to this point in Example 2.11.

**Example 2.5** (Random sample.) Let $Y = (Y_1, Y_2, Y_3, \ldots)$ be random sample from a fixed community $\pi$ where $Y_i$ is the number of times the $i$-th ranked species is represented in the sample. Then $Y/N$ is a random community whose abundance vector has a rescaled multinomial distribution. Note that $E[Y/N] = \pi^\ast$.

**Example 2.6** (Random permutation.) Let $\nu$ be obtained from the random community $\pi$ by subjecting the components of $\pi$ to a permutation. In general, the permutation may be random and its distribution may depend on the realized value of $\pi$. Species identity may be lost after a random permutation since, for example, $\nu_1$ may reference different species in different realizations. But species identity is unimportant for questions of diversity so that $\pi$ and $\nu$ may be regarded as equivalent.

The issue of permuting $\pi$ is essentially trivial for fixed communities and was glossed over in Section 1. But for random communities it may be that some carefully chosen permutation of $\pi$ has a tractable distribution. Also a sequence of permuted communities may have a limiting distribution even though the original sequence does not.

Two types of random permutations are quite useful. The ranked permutation $\nu = \pi^\ast$ arranges the components of $\pi$ in descending order and is canonical in the sense that $\pi^\ast$ and $\pi^\ast$ are equal in distribution whenever $\pi'$ is a random permutation of $\pi$.

The size-biased permutation $\nu = \pi^\dagger$ is obtained as follows: Randomly select an individual from the community and put $\pi^\dagger_1 = \pi_{i(1)}$ where $i(1)$ is the label of the species to which the selected individual belongs. Now remove species $i(1)$ from the community and randomly select a second individual. Put $\pi^\dagger_2 = \pi_{i(2)}$ where this second individual belongs to species $i(2)$. Now remove species $i(1)$ and $i(2)$ and select a third individual, etc.

The size-biased permutation arises naturally in the problem of “heaps” which has been considered by Kingman (1975). A number of items labelled $i = 1, 2, \ldots, N$ are stored in a heap and are demanded from time to time. After being used, the item is replaced on the top of the heap. The successive arrangements of items in the heap form an irreducible aperiodic Markov
chain with \( N \) states. The equilibrium distribution is easily seen to be the
distribution of \( \pi^* \) where \( \pi_i \) is the proportional demand for item \( i \).

As a numerical example of size-biased permutation, take \( \pi \) to be the fixed
community \((.6, .3, .1)\). The probability distribution of the random community
\( \pi^* \) is given in Table 2. Notice that \( \pi^* \) arranges the components of \( \pi \) so that
large abundances are likely, but not certain, to appear early in the list. The
next theorem makes this precise.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( P(\pi^* = x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(.6, .3, .1)</td>
<td>.480</td>
</tr>
<tr>
<td>(.6, .1, .3)</td>
<td>.160</td>
</tr>
<tr>
<td>(.3, .6, .1)</td>
<td>.267</td>
</tr>
<tr>
<td>(.3, .1, .6)</td>
<td>.043</td>
</tr>
<tr>
<td>(.1, .6, .3)</td>
<td>.087</td>
</tr>
<tr>
<td>(.1, .3, .6)</td>
<td>.033</td>
</tr>
</tbody>
</table>

**Theorem 2.1:** Let \( \pi^* \) be the size-biased permutation of the random
community \( \pi \). Then \( \pi^* \) is stochastically greater than \( \pi_{i,1}^* \), \( i = 1, 2, 3, \ldots \). Consequently, \( E[\pi^*_1] > E[\pi^*_2] > E[\pi^*_3] > \ldots \).

**Example 2.7:** (Symmetric Dirichlet.) Let \( \pi \) have a symmetric Dirichlet
distribution with \( s \) components and with index \( k > 0 \). Write \( \pi \sim D(s, k) \).
With probability one, every realization has exactly \( s \) species. Recall that
\( \pi \) can be obtained as \( (\lambda_1, \lambda_2, \ldots, \lambda_s)/\Sigma \lambda_i \) where \( \lambda_1, \lambda_2, \ldots, \lambda_s \) are iid Gamma
random variables with common index \( k \) and with common, but arbitrary,
scale parameter. A useful fact is that the fractions are jointly independent
of the total, i.e., \( \pi \) is independent of \( \lambda_1 + \lambda_2 + \ldots + \lambda_s \). The size-biased permutation of the Dirichlet is quite manageable
and a number of its properties are discussed below. The ranked permutation, on the other hand, is usually
intractable; the Dirichlet with \( k = 1 \) is an exception and is equal in distribution
to MacArthur's (1957) broken stick model.

**Preemption and the Limiting Dirichlet**

Given the random community \( \pi \), define the residual fractions \( Q_1, Q_2, \ldots, Q_s \)
as
\[
Q_1 = \pi_1 \\
Q_2 = \pi_2/(1-\pi_1) \\
Q_3 = \pi_3/(1-\pi_1-\pi_2) \\
\vdots \\
Q_s = \pi_s/(1-\pi_1-\pi_2-\ldots-\pi_{s-1}) = 1.
\]
The residual fractions specify a sequential resource allocation scheme: Let
the total resource be represented by the unit interval as in MacArthur's model.
A first species preempts a fraction $Q_1$ of the total, then a second species preempts
a fraction $Q_2$ of the residual $1 - \pi_1$, then a third species preempts a fraction
$Q_3$ of the new residual $1 - \pi_1 - \pi_2$, etc. The random community is called a
residual allocation model (or a preemption model) when the residual fractions
are independently distributed with $P(0 < Q_i < 1) = 1$, $i = 1, 2, 3, \ldots, s - 1,$
and $P(Q_s = 1) = 1$. If the number of species is infinite we also require that
\[ P(\lim_{n \to \infty} (1 - \pi_1 - \pi_2 - \ldots - \pi_n) = 0) = 1. \]
\[ \quad \text{(2.1)} \]

Theorem 2.2: Equation (2.1) holds when $Q_1, Q_2, Q_3, \ldots$ are independent
and identically distributed.

It is well-known that the Dirichlet with parameters $s$ and $k$ is a residual
allocation model with $Q_1 \sim \text{Beta}_s[k, (s - i)k]$. The size-biased permutation
of the Dirichlet is also a residual allocation model but with $Q_1 \sim \text{Beta}_s[k + 1,
(s - i)k]$. It would be interesting to characterize the class of residual allocation
models that possess this type of invariance to size-biased permutation.

Example 2.8: (Engen’s model.) This is the infinite residual allocation
model $\pi$ whose residual fractions are iid with common distribution $\text{Beta}_s
(1, \alpha)$ where $\alpha > 0$. The expectations $E[\pi_1], E[\pi_2], E[\pi_3], \ldots$ form a geometric
series with ratio $\theta = \alpha/(\alpha + 1)$ and odds ratio $\theta/(1 - \theta) = \alpha$. The model is
thus a randomized form of the geometric series. Engen (1975) has associated
this model with Fisher’s limiting Dirichlet which we discuss next.

Example 2.9: (Limiting Dirichlet.) For the Dirichlet model, Fisher
(1943) introduced the product $\alpha = sk$ as a diversity measure. For fixed $k$,
$\alpha$ is proportional to the number of species and may be regarded as a species
richness parameter. As justification for the use of $\alpha$, Kempton and Taylor
(1974) point out that, for most data sets, it is not possible to efficiently estimate
$s$ and $k$ separately although the product $\alpha = sk$ can be efficiently estimated.
Fisher found that $s$ was usually very large and $k$ very small which led him to
idealize the model by taking the formal limit as $s \to \infty$ and $k \to 0$ with the
product $\alpha = sk$ held fixed. The sampling distribution considered by Fisher
did converge (to the log series distribution) permitting the development of inferential techniques. But the underlying population model does not
converge.
The reason the Dirichlet fails to converge becomes clear once we note that there is no community $\pi$ with infinitely many species and whose components $\pi_i$ are identically distributed. As Kingman (1975) points out, such a community would have to satisfy the incompatible requirements

$$E[\pi_1] = E[\pi_2] = E[\pi_3] = \ldots,$$

$$\Sigma E[\pi_i] = 1.$$

This difficulty can be avoided by subjecting the Dirichlet community to an appropriate canonical random permutation before taking the limit. For example, results of Kingman (1975) imply that the ranked permutation $\pi^*$ does converge in distribution under Fisher's limiting scheme. Unfortunately, the limiting distribution of $\pi^*$ is quite intractable.

On the other hand, the size-biased permutation converges to Engen's model with parameter $\alpha$. Furthermore, the ranked permutation of Engen's model is equal in distribution to Kingman's limit and the size-biased permutation of Kingman's limit equals Engen's model. This has two important implications. First, the two models are permutations of one another and are merely alternative descriptions of the same underlying community structure. The choice of which to use can be made on grounds of mathematical convenience and Engen's model clearly wins out. Second, Engen's model is itself invariant under size-biased permutation. This is because Engen's model is already the size-biased permutation of Kingman's limit and $\pi^*$ equals $\pi^*$ in distribution. We have used this invariance to give a characterization of Engen's model (unpublished) :

Theorem 2.3: Engen's model is the only infinite residual allocation model whose residual fractions are iid and which is invariant under size-biased permutation.

As an application of size-biased invariance, consider the problem of calculating $E[\Delta(\pi)]$ where $\pi$ is a random community and where $\Delta(\pi) = \Sigma \pi_i R(\pi_i)$ is a dichotomous index with rarity measure $R$. Now $\Sigma \pi_i R(\pi_i)$ is the conditional expectation of $R(\pi^*)$ so that

$$E[\Delta(\pi)] = E[R(\pi^*)].$$  \hspace{1cm} (2.2)

When $\pi$ is Engen's model with parameter $\alpha$, size-biased invariance implies that $\pi^* \sim \pi_1 \sim \text{Beta}(1, \alpha)$ and (2.2) becomes

$$E[\Delta(\pi)] = \alpha \int_0^1 R(x)(1-x)^{\alpha-1} \, dx.$$  \hspace{1cm} (2.3)
This may be compared with equation (20) of Kingman (1975). Calculation of the $E\Delta$ measures is thus reduced to the evaluation of an integral, but an implication of (2.3) is of more interest to us. Since the distribution Beta $(1, \alpha)$ is stochastically decreasing in $\alpha$, we have,

**Theorem 2.4:** Assume that $R$ is decreasing on the interval $(0, 1]$ (Criterion C1). Then $E[\Delta(\pi_1)]$ is an increasing function of $\alpha$ when $\pi$ is Engen's model with parameter $\alpha$.

Recall that Criterion C1 ensures that introducing a species—but not necessarily transferring abundance—increases the diversity index. Theorem 2.3 clearly suggests that $\alpha$ should be looked upon as a species richness parameter and that two limiting Dirichlets differ only in their "number" of species. The next example is intended to lend some precision to this point of view.

**Example 2.10:** (Random deletion of species.) Let $0 < p < 1$ and suppose $\nu$ is a random community with $P(0 < \nu_i < 1) = 1$ for $i = 1, 2, 3, \ldots$. Further suppose $\epsilon_1, \epsilon_2, \epsilon_3, \ldots$ are independent Bernoulli random variables with $P(\epsilon_i = 0) = 1-p$ and $P(\epsilon_i = 1) = p$. Put $\pi_i = \nu_\epsilon_i / \sum_\nu \epsilon_i$. Then $\pi = \pi(y, \pi_2, \pi_3, \ldots)$ is a random community obtained by screening $\nu$. Call $\epsilon_1, \epsilon_2, \epsilon_3, \ldots$ the selection variables. Defined in this way, $\pi$ has infinitely many components equal to zero. If desired, the zero components can be omitted by the following device: Let $X(1), X(2), X(3), \ldots$ be independent geometric random variables with $P(X(i) = x) = p(1-p)^{x-1}, x = 1, 2, 3, \ldots$. Put $\tau_i = \nu_{X(1)+\cdots+\nu_{X(2)}+\cdots+\nu_{X(i)}}$ and $\pi_i = \tau_i / \tau$ where $\tau = \tau_1 + \tau_2 + \tau_3 + \cdots$. The following theorem may be proved using the Gamma process.

**Theorem 2.5:** Let $\nu$ be Engen's model with parameter $\alpha$. Then $\alpha$) $\tau = \Sigma \tau_i$ has a Beta $(\alpha, (1-p)\alpha)$ distribution and is independent of $\tau = (\tau_1, \tau_2, \tau_3, \ldots) / \tau$, $\beta$) $\pi$ is equal in distribution to Engen's model with parameter $\alpha$.

**Remark:** Omitting the zeros is essential if $\tau$ is to be independent of $\pi$.

The theorem also holds for Kingman's version of the limiting Dirichlet.

**Dominance**

It does not seem to be widely recognized that the geometric series may be generated by two essentially different mechanisms, which we call preemption and dominance. Preemption supposes that the first species preempts a fraction $Q_1$ of the available resource, the second a fraction $Q_2$ of the remainder, etc. In dominance, the first species receives an amount of resource equal to $P_1$, the second species receives a fraction $P_2$ as much as the first, the third species receives a fraction $P_3$ as much as the second, etc. Both preemption and dominance result in a geometric series of species abundances.
when \( P_i = 1 - Q_i \) are all equal to some fixed number. The stochastic analogues are obtained by letting \( P_i = 1 - Q_i \) be iid random variables. Here preemption and dominance are quite different. Due to randomness, the successive abundances assigned by the preemption mechanism need not be decreasing. Hence, the expected ranked abundance vector may not be a geometric series. In dominance, however, the successive absolute abundances are decreasing and have geometric expectations.

When \( Q_i \sim \text{Beta}_1 (1, \alpha) \), the preemption mechanism gives the size-biased version of the limiting Dirichlet. This has led Engen (1975) to recommend the geometric series as a fixed version of the limiting Dirichlet. For the reasons indicated in Examples 2.4 and 2.11, we do not think the geometric series is an appropriate choice. More appropriate is Watterson's (1974) exponential integral model. At a later date, we intend to examine in greater depth the matter of replacing a random community with some typical fixed version. In particular, an extension of Watterson's (1974) method shows that the random community corresponding to the geometric series is the dominance mechanism with \( P_i = 1 - Q_i \sim \text{Beta}_1 (\alpha, 1) \).

This model can also be obtained via a Fisher-like limit. Consider the classical framework of \( s \) species having iid abundances \( \lambda_1, \lambda_2, \ldots, \lambda_s \). Motivated by the dominance mechanism, we would like the expected descending order statistics to form a geometric series, say \( E[\lambda_{(s-t)}^+] = \theta^t, t = 1, 2, \ldots, s, 0 < \theta < 1 \). Kadane (1971) has shown that, for each value of \( s \), there exists such a probability distribution. He did not explicitly give the distribution, but, using his methods, the probability density function can be obtained as

\[
\theta(1-\theta)^{-t(s-1)}u^{1-1/(s+1)}, \quad \theta^{s+1} < u < 1. \quad \ldots \ (2.4)
\]

Note that the functional form of this density as well as the range, depends explicitly on \( s \). Letting \( s \to \infty \) and ignoring the normalizing constant, (2.4) converges to the improper density \( f(u) = 1/u, 0 < u < 1 \). Now \( f(u) \) may be taken as the Levy density of a subordinator process. Let the process be observed during the time interval \( 0 < t < \alpha \) and the jumps arranged in descending order. It may be shown that these ordered jumps are equal in distribution to the abundances generated by the dominance mechanism with \( P_i \sim \text{Beta}_1 (\alpha, 1) \).

Pielou (1975, p. 21) has suggested that the expected descending order statistics \( E[\lambda^+] \) are approximately a geometric series when \( \lambda_1, \lambda_2, \ldots, \lambda_s \) have probability density function

\[
(-\log(c))^{-1/u}, \quad 0 < c < u < 1. \quad \ldots \ (2.5)
\]
Note that (2.5) compares favorably with (2.4) for large \( s \) provided the truncation parameter \( c \) is approximately \( \theta^{s+1} \). The exact expected order statistics from (2.5) are

\[
E[\lambda^*_i] = c M(s + 1 - i, s + 1, -\log(c)), \quad i = 1, 2, ..., s,
\]

where \( M(\cdot, \cdot, \cdot) \) is Kummer's confluent hypergeometric function. Using (13.4.1) of Abramowitz and Stegun (1965), one obtains

\[
i E[\lambda^*_i] + (s + 1 - 2i - \log(c)) E[\lambda^*_i] - (s + 1 - i) E[\lambda^*_{i-1}] = 0.
\]

From this, it follows that the expected order statistics converge to a geometric series with ratio \( \alpha/(\gamma + 1) \) when \( s \to \infty \) and \( c \to 0 \) such that \( \log(c)/s \to -1/\alpha \).

The preceding calculations are of limited practical interest since, as Pielou (1975, p. 23) points out, the dominance model does not often fit empirical data and then only when the number of species is small. This is consistent with our own findings that the dominance model has a sampling distribution whose tail is much shorter than the log series. Empirical data usually requires a tail at least as long as the log series.

**Stochastic Diversity**

This section extends the intrinsic diversity ordering to random communities. As was the case for fixed communities the ordering is only partial and two random communities need not be comparable. However, some comparability relations are established for Dirichlet communities and these shed further light on the role of Fisher's "alphas" as a diversity parameter.

The intrinsic diversity ordering has several stochastic analogues corresponding to the various parts of Theorem 1.3. Let \( \pi \) and \( \nu \) be random communities and define \( \nu \) to be **stochastically more diverse** than \( \pi \) in the sense (SD2, SD3, SD4, SD5) if:

- **SD2**: \( E[\Delta(\nu)] \geq E[\Delta(\pi)] \) whenever \( \Delta \) satisfies Criterion C2.
- **SD3**: \( E[\Delta(\nu)] \geq E[\Delta(\pi)] \) whenever \( \Delta \) satisfies Criterion C3.
- **SD4**: For every positive integer \( k \), the random variable \( \sum_{i \leq k} \pi^*_i \) is stochastically greater than \( \sum_{i \leq k} \nu^*_i \).
- **SD5**: The fixed community \( E[\nu^*] \) is intrinsically more diverse than the fixed community \( E[\pi^*] \).

Stochastic diversity in the sense SD2 is the dual of stochastic majorization as defined by Nevis, Proschan and Sethuraman (1977). In the same paper,
these authors have shown (Theorem 2.2) that the following statements are equivalent:

i) $\nu$ is stochastically more diverse than $\pi$ in the sense SD2.

ii) The random variable $\Delta(\nu)$ is stochastically greater than $\Delta(\pi)$ whenever $\Delta$ satisfies Criterion C2.

iii) $P(\nu \in A) \geq P(\pi \in A)$ whenever $A$ is a Schur convex set.

Statement (iii) is a direct requirement on the probability distributions of $\pi$ and $\nu$ and is similar to the definition of stochastic ordering of random variables. For this reason SD2 is the most elegant and, perhaps, the most appropriate formulation of stochastic diversity. However, SD2 is a very stringent requirement. A few of the results established below make use of Jensen's inequality and apply to the definition SD3.

First the logical relations that exist among the alternative formulations of stochastic diversity will be given.

Theorem 2.6: Each of the following implications is true:

- SD3
- SD2
- SD4
- SD5

In general, none of these implications can be reversed; however, for fixed communities the four definitions are equivalent.

Example 2.11: This is the first application of Jensen's inequality. Let $\pi$ be a random community and $\Delta$ an index satisfying Criterion C3. Then $E[\Delta(\pi)] = E[\Delta(\pi^*)] \leq \Delta(E[\pi^*])$ which shows that:

a) $E[\pi^*]$ is stochastically more diverse than $\pi$ in the sense SD3.

Now let $\nu$ be any random permutation of $\pi$ for which $E[\nu]$ is ranked, i.e. $E[\nu_1] > E[\nu_2] > E[\nu_3] > \ldots$; for example, $\nu$ might be the size-biased permutation. Clearly $\pi_1^* + \pi_2^* + \ldots + \pi_k^* > \nu_1 + \nu_2 + \ldots + \nu_k$ for any positive integer $k$. Taking expectations shows that:

b) $E[\nu]$ is intrinsically more diverse than $E[\pi^*]$, and
c) $E[\pi^*]$ is intrinsically more diverse than $E[\pi^*]$.

Taken together, these relations indicate that fixing a random community by the method of expectations will exaggerate the community's diversity.
Most of the results in Nevius, Proshcan and Sethuraman (1977) are stated under an assumption that certain parameters \( \lambda \) and \( \lambda_i \) are positive. In the present terminology, this requires that both communities have the same number of species which is too restrictive for applications to diversity. Fortunately, the assumption can be relaxed to the requirement that \( \lambda \) and \( \lambda_i \) be nonnegative. Simply note that the proof of the preservation theorem as given by Proshcan and Sethuraman (1977) remains valid for \( \lambda = 0 \) provided \( \phi(x) \) is interpreted as the Dirac delta function with unit mass at \( x = 0 \). With this extension, Application 4.2(a) of Nevius, Proshcan and Sethuraman (1977) implies that random samples reflect the intrinsic diversity ordering in the following sense:

**Theorem 2.7:** Let \( Y \) and \( Y' \) be random samples of size \( N \) from the fixed communities \( \pi \) and \( \pi' \) as in Example 2.5. Then \( Y'/N \) is stochastically more diverse than \( Y/N \) in the sense SD2 when \( \pi' \) is intrinsically more diverse than \( \pi \).

The rest of this section is devoted to diversity comparisons of Dirichlet communities. The Dirichlet model has the two parameters \( s \) and \( k \) as well as the derived parameter \( \alpha = sk \) and any diversity relations will have to involve two of these parameters. Necessary and sufficient conditions for the comparability of two Dirichlet communities have not been obtained, but the next theorem is a close approximation.

**Theorem 2.8:** Let \( \pi \) and \( \pi' \) be Dirichlet communities with parameters \( (s, k) \) and \( (s', k') \), respectively. In order that \( \pi' \) be stochastically more diverse than \( \pi \) in the sense SD2: (a) the pair of conditions \( s' \geq s \) and \( (s' - 1) k' \geq (s - 1) k \) are necessary, and (b) the pair of conditions \( s' \geq s \) and \( s' k' \geq sk \) are sufficient but not necessary.

The necessary conditions of the theorem are obtained by requiring that \( E[\Delta(\pi')] \geq E[\Delta(\pi)] \) for appropriate choices of the index \( \Delta \). The choices should include indices sensitive to rare species as well as indices sensitive to abundant species. The choice of \( \Delta_{-1} = s - 1 \) establishes the necessity of \( s' \geq s \). The other necessary condition is obtained by taking \( \Delta = \Delta_{s} \) and letting \( \beta \to \infty \).

The proof of (b) is based in part on Application 4.2(d) of Nevius, Proshcan and Sethuraman (1977).

It seems likely that the conditions given in Theorem 2.18(a) are both necessary and sufficient. This has been verified for small values of \( s \) and \( s' \). In any event, the conditions of part (a) and of part (b) are numerically indistinguishable for realistic parameter values (\( s \) large and \( k \) small).

We conclude the paper by showing that Fisher's \( \alpha \) is a complete diversity parameter for the limiting Dirichlet, at least in the sense SD3. The proof is
an interesting application of random deletion of species and will be presented in detail.

Theorem 2.9: Let \( \nu \) and \( \pi \) be limiting Dirichlet models with parameters \( \alpha \) and \( \alpha_0 \), respectively. Then \( \nu \) is stochastically more diverse than \( \pi \) in the sense \( SD3 \iff \alpha \geq \alpha_0 \).

Proof: (\( \iff \)). Since the Simpson index satisfies Criterion C3, we must have \( E[\Delta(\nu)] \geq E[\Delta(\pi)] \) or, by (2.3), \( \alpha/(\alpha+1) \geq \alpha_0/(\alpha_0+1) \). This implies that \( \alpha \geq \alpha_0 \). (\( \iff \)). Without loss of generality, assume \( \alpha > \alpha_0 \) and write \( \alpha_0 = p\alpha \) where \( 0 < p < 1 \). Now \( \pi \) may be obtained from \( \nu \) by random deletion of species with zeros omitted, as in Theorem 2.4. Let \( \epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, \ldots) \) be the selection variables (see Example 2.10). The sample space \( W \) of \( \epsilon \) consists of all infinite strings of zeros and ones. Let \( d\mu(\epsilon) \) be the probability measure on this space which specifies the distribution of \( \epsilon \). For a given realization of \( \nu \), let \( \nu \times \epsilon \) be the vector whose \( i \)-th component is \( \nu_i \epsilon_i \) and put \( S(\nu, \epsilon) = \Sigma_i \epsilon_i \). Then \( \nu \times \epsilon/S(\nu, \epsilon) \) is the same as \( \pi \) but without the zeros omitted. Since \( \int \epsilon \ d\mu(\epsilon) = p \), one easily checks that, for a given realization of \( \nu \), \( S(\nu, \epsilon) \ d\mu(\epsilon)/p \) is a probability measure on \( W \) and

\[
\nu = \int_W (\nu \times \epsilon/S(\nu, \epsilon)) S(\nu, \epsilon) d\mu(\epsilon)/p.
\]

Now let \( \Delta \) satisfy Criterion C3. By Jensen's inequality (2.8) gives

\[
\Delta(\nu) \geq \int \Delta(\nu \times \epsilon/S(\nu, \epsilon)) S(\nu, \epsilon) d\mu(\epsilon)/p
\]

\[
= E[\Delta(\nu \times \epsilon/S(\nu, \epsilon)) S(\nu, \epsilon)/\nu]/p.
\]

Take expectations with respect to \( \nu \) to find that

\[
E[\Delta(\nu)] \geq E[\Delta(\nu \times \epsilon/S(\nu, \epsilon)) S(\nu, \epsilon)]/p.
\]

But \( \Delta(\nu \times \epsilon/S(\nu, \epsilon)) = \Delta(\pi) \) since omitting the zero components does not change the value of \( \Delta \). By Theorem 2.4, \( \pi \) is independent of \( S(\nu, \epsilon) \) so that

\[
E[\Delta(\nu)] \geq E[\Delta(\pi)] E[S(\nu, \epsilon)]/p = E[\Delta(\pi)].
\]

Open Problem: Is Theorem 2.8 true in the sense SD2 ?

References


Key words


Abstract

The paper has two sections. The first summarizes some aspects of the authors' earlier work on diversity and reports a new unifying formulation of the measurement of species rarity. Various indices, including the Hurlbert-Smith index, are then obtained as special cases. The second section is concerned with random communities and has, as its central theme, the justification of Fisher's "alpha" as a diversity measure associated with the symmetric Dirichlet model.

Résumé

La présente communication comporte deux parties. La première présente un survol de travaux antérieurs de l'auteur et on y rapporte une nouvelle formulation unifiée de la mesure de la raréité d'une espèce. On obtient alors comme cas particuliers divers indices, y compris celui de Hurlbert-Smith. La seconde partie porte sur les communautés à composition aléatoire et a pour thème principal la justification de l'"alpha" de Fisher en tant que mesure de la diversité, mesure liée au modèle symétrique de Dirichlet.