UMAP

Modules in Undergraduate Mathematics and Its Applications

Published in cooperation with the Society for Industrial and Applied Mathematics, the Mathematical Association of America, the National Council of Teachers of Mathematics, the American Mathematical Association of Two-Year Colleges, The Institute of Management Sciences, and the American Statistical Association.

COMAP

Module 705

Information Theory and Biological Diversity

Steven Kolmes
Kevin Mitchell

Applications of Calculus to Biology

INTERMODULAR DESCRIPTION SHEET:	UMAP Unit 705
TITLE:	Information Theory and Biological Diversity
AUTHOR:	Steven Kolmes Dept. of Biology University of Portland Portland, OR 97203 and Kevin Mitchell Dept. of Mathematics and Computer Science Hobart and William Smith Colleges Geneva, NY 14456
MATHEMATICAL FIELD:	Calculus
APPLICATION FIELD:	Biology, ecology
TARGET AUDIENCE:	Students in first- or second-semester calculus course or in a biological modeling course
ABSTRACT:	Discusses and derives the key properties of one measure of diversity, the entropy function, and illustrates its use by ecologists and animal behaviorists.
PREREQUISITES:	1. Familiarity with the Mean Value Theorem. 2. Familiarity with logarithms and their derivatives. 3. Ability to use the first and second derivatives as aids in graphing functions. 4. L'Hôpital's rule (used once).
OUTPUT SKILLS:	1. Familiarity with one common measure of diversity: $-\sum_{i=1}^{n} p_i \log_2 p_i$. 2. Know how to apply this measure to ecological problems. 3. Greater appreciation of the Mean Value Theorem.
SUGGESTED ADDITIONAL RESOURCE:	Scientific calculator

Information Theory and Biological Diversity

Steven Kolmes

Department of Biology
University of Portland
Portland, OR 97203

Kevin Mitchell

Department of Mathematics and Computer Science
Hobart and William Smith Colleges
Geneva, NY 14456

Table of Contents

MODULES AND MONOGRAPHS IN UNDERGRADUATE
MATHEMATICS AND ITS APPLICATIONS (UMAP) PROJECT

The goal of UMAP is to develop, through a community of users and developers, a system of instructional modules in undergraduate mathematics and its applications, to be used to supplement existing courses and from which complete courses may eventually be built.

The Project was guided by a National Advisory Board of mathematicians, scientists, and educators. UMAP was funded by a grant from the National Science Foundation and now is supported by the Consortium for Mathematics and Its Applications (COMAP), Inc., a nonprofit corporation engaged in research and development in mathematics education.

Paul J. Campbell   Editor
Solomon Garfunkel   Executive Director, COMAP

1.   Introduction

This Module is concerned with making the notion of "diversity" precise. In commonplace usage, the term is synonymous with "variety" and is simply an indication of the number of different things that are present. For example, we often speak of a "diversity of opinions." While simply counting the number of different types of opinions on a subject can give a rough idea of the "diversity of opinions," the numbers of people holding the various opinions must be taken into account to get a true sense of the diversity. For example, a situation in which there are 99 people with one opinion and 1 person holding a different opinion is different from a situation in which 50 people have one opinion and the other 50 have another, even though the number of opinions (2) is the same in both cases.

Biologists use a mathematical concept called *information* to make precise calculations about entities that we will come to know as *first-order diversity*, H_1, and *divergence from equiprobability*, D_1. We will explore how information theory operates and examine several biological applications of these concepts. In particular, these mathematical entities have proven to be quite useful to ecologists and animal behaviorists. This single set of mathematical formulations, originally intended for use in designing communications systems, has an unusually broad range of applications.

2.   Building Intuition

The following example illuminates the properties that any measure of diversity should have. Assume that you have planted a garden with four types of flowers in equal numbers. Over the course of a growing season, you return to the garden several times to chart the progress of the plants.

In early May, you observe that all of the first three types of flowers have begun to grow. However, only a few of the fourth type of flower are growing. You carefully chart the number of flowers of each flower type as a proportion of the total number of flowers in the garden. Data describing the diversity of plants present would look like **Table 1**.

Table 1.

The garden in May: very few plants of type 4 present.

Flower Type	1	2	3	4
Proportion	5/16	5/16	5/16	1/16

In June, you find that all of the type 4 plants have now sprouted. Each of the four types of plants is now present in equal numbers in the garden. (A situation in which equal proportions of each category are present is a state of *equiprobability*, though "equiproportionality" might be more accurate.) Data

describing the diversity of plants would now look like **Table 2**. Notice that the number of types of plants does not change from May to June. However, in May the first three types of flowers dominated the garden, and the fourth type was barely present. By June, all four types of plants are present in equal numbers; none is dominant. Because of this, the diversity of the flower garden increased from May to June.

Table 2.

The garden in June: equiprobability.

Flower Type	1	2	3	4
Proportion	1/4	1/4	1/4	1/4

By August the heat of summer has caused all the flowers of type 2 and half of those of types 1 and 3 to wither and die, though all of the flowers of type 4 are still alive (**Table 3**). The diversity in the garden has decreased from June, because the number of types of flowers present has decreased and one type of flower dominates the garden. It is even less diverse than in May, when three types of flowers were present in equal numbers and a fourth type was present in low numbers.

To assure yourself that the garden in August is less diverse than in May, imagine trying to guess which type of flower you would encounter next in a walk through the garden. In August there is one very common type. By always guessing "type 4" you would be right on average half the time. In May there were three equally common flower types and an additional rare type. Guessing that your next encounter would be with one of the more common types of flower would on average lead to success only five-sixteenths of the time. The greater the diversity of the system, the harder it is to predict what will be encountered next.

Table 3.

The garden in August: low diversity.

Flower Type	1	2	3	4
Proportion	1/4	0	1/4	1/2

By late September, flower types 1 and 3 are now gone because of an early frost; only flower type 4 remains (**Table 4**). The garden is not only less diverse than in August, the garden has no diversity of plant life at all. Every plant is of a single type.

What are we to make of this example? Any method we propose to measure diversity needs to reflect the observations that we have just made. Both the number of categories present and the proportions in each category contribute to the overall diversity. The comparison of the May and June gardens indicates that diversity should be maximized when all categories are present in equal

Table 4.

The garden in September: no diversity at all.

Flower Type	1	2	3	4
Proportion	0	0	0	1

proportions and no single category predominates. Secondly, when a system is in such a state of equiprobability, the more categories the system has, the more diverse it will be. (A garden with eight types of flowers present in equal proportions should be more diverse than the June garden with only four types present in equal proportions.) Finally, when all the items fall into a single category, as in the September garden, the measure of diversity should be 0. To summarize, any measure of diversity should satisfy the following three conditions.

Condition 1. *The measure of diversity should be 0 when one of the categories has proportional representation 1 and the rest are represented at a proportion of 0 (not seen).*

Condition 2. *If there are n possible categories, the diversity measure should be maximized when all categories are observed equally, that is, $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$, where p_i is the proportion of the ith category.*

Condition 3. *If $m > n$, then a state of equiprobability (every category observed equally) for a system with m categories should have a higher diversity than a state of equiprobability for a system with only n possible categories.*

3.   Diversity Defined

3.1   Proportions

The biological applications discussed in this Module are based on reports of field observations. Typically such data are given as an array of proportions of times that particular types of observations were made, much like the data in **Tables 1–4**. Each of the experiments reported will have only a finite number of categories of objects.

Suppose that in a particular experiment exactly n distinct categories were observed, e_1, e_2, ..., e_n, and that the respective proportions of time they were observed were p_1, p_2, ..., p_n. Then these proportions must satisfy two basic properties.

Property 1. *Each of the numbers p_1, p_2, ..., p_n lies between 0 and 1.*

A proportion of 1 means that all of the observations belong to a single category. A proportion of 0 indicates that no objects belong to that category.

(Such a category might still be included in the list if it occurred in some other phase of the experiment.)

Property 2. $\displaystyle\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1.$

Property 2 says that the list of events is complete; the proportions of the events e_1, e_2, ..., e_n must sum to 100% of the observations. Notice that the data in **Tables 1–4** satisfy both of these elementary properties.

3.2 The Measure of Diversity

The proportions of different categories of objects in the environment clearly play a role in our intuitive notion of diversity, as illustrated in the garden example above. In that example we were able to rank the relative diversities of the gardens in each month. The June garden had the highest diversity, then May, August, and finally September. However, such a ranking does not permit us to quantify the differences in diversity among the gardens. And such a ranking might be difficult to form when there were 10, 20, or even hundreds of different species of flowers.

How do we make all of this precise? Mathematicians have found a function that satisfies the three conditions listed in Section 2 and also satisfies a number of other important conditions.

Definition 1. *Assume that there are n possible categories in an experiment and that their proportions are p_1, ..., p_n. Then the **measure of diversity**[1] for this system is*

$$H_1 = -\sum_{i=1}^{n} p_i \log_2 p_i = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \cdots + p_n \log_2 p_n).$$

The units of measurement are called *bits*. (Since $\log_2 0$ is not defined, if $p_i = 0$ we adopt the convention that the expression $p_i \log_2 p_i = 0 \log_2 0$ is also 0.)

At this point you probably feel uncertain about the function H_1. As strange as it looks at first glance, this function has become extraordinarily useful not only in mathematics and engineering (especially communications), but also in many of the natural and social sciences. It is worth struggling with because it represents a relatively simple way of quantifying the extremely abstract concept of diversity.

---

[1]In other contexts, this function is also known as the *measure of uncertainty*, *measure of disorder*, or the *entropy* of the system.

3.3   Playing with H_1

To become comfortable with this measure of diversity, we will work out various examples. Let us start by evaluating the different levels of diversity for the flower garden example in Section 2 in each of the four months. In each month there were four possible categories, whose proportions were listed in **Tables 1–4**.

At the first stage, in May (**Table 1**), we have

$$
\begin{aligned}
H_1 &= -\sum_{i=1}^{4} p_i \log_2 p_i \\
&= -\left(\tfrac{5}{16} \log_2 \tfrac{5}{16} + \tfrac{5}{16} \log_2 \tfrac{5}{16} + \tfrac{5}{16} \log_2 \tfrac{5}{16} + \tfrac{1}{16} \log_2 \tfrac{1}{16} \right) \\
&= -\tfrac{15}{16} \log_2 \tfrac{5}{16} - \tfrac{1}{16} \log_2 \tfrac{1}{16} \\
&= \tfrac{15}{16} \log_2 \tfrac{16}{5} + \tfrac{1}{16} \log_2 16 \\
&\approx 1.823 \,.
\end{aligned}
$$

In June (**Table 2**) we have equiprobability, so

$$
\begin{aligned}
H_1 &= -\sum_{i=1}^{4} p_i \log_2 p_i \\
&= -\left(\tfrac{1}{4} \log_2 \tfrac{1}{4} + \tfrac{1}{4} \log_2 \tfrac{1}{4} + \tfrac{1}{4} \log_2 \tfrac{1}{4} + \tfrac{1}{4} \log_2 \tfrac{1}{4} \right) \\
&= -4 \left(\tfrac{1}{4} \log_2 \tfrac{1}{4} \right) \\
&= -\log_2 \tfrac{1}{4} \\
&= \log_2 4 \\
&= 2 \,.
\end{aligned}
$$

Notice that the diversity has increased from May to June, as we expected.

With fewer categories present in August (**Table 3**), the diversity is reduced to

$$
\begin{aligned}
H_1 &= -\sum_{i=1}^{4} p_i \log_2 p_i \\
&= -\left(\tfrac{1}{4} \log_2 \tfrac{1}{4} + 0 \log_2 0 + \tfrac{1}{4} \log_2 \tfrac{1}{4} + \tfrac{1}{2} \log_2 \tfrac{1}{2} \right) \\
&= -\tfrac{1}{2} \log_2 \tfrac{1}{4} - \tfrac{1}{2} \log_2 \tfrac{1}{2} \\
&= \tfrac{1}{2} \log_2 4 + \tfrac{1}{2} \log_2 2 \\
&= 1.5 \,.
\end{aligned}
$$

The diversity in August is lower than in both May and June, and the difference between May and August is larger than the difference between May and June. This reflects the reduction in the number of flower types present in August, as well as the higher proportion of flower type 4.

In September (**Table 4**), as we know, there is no diversity in the garden. Indeed,

$$
\begin{aligned}
H_1 &= -\sum_{i=1}^{4} p_i \log_2 p_i \\
&= -(0\log_2 0 + 0\log_2 0 + 0\log_2 0 + 1\log_2 1) \\
&= -\log_2 1 \\
&= 0.
\end{aligned}
$$

In fact, in a completely analogous fashion, H_1 will *always* be 0 whenever one of the proportions is 1 and the rest are 0, since $\log_2 1 = 0$. That is, whenever there is complete certainty as to the outcome, $H_1 = 0$. Thus H_1 satisfies the first condition we placed on a measure of diversity, at the end of Section 2.

There are other general observations we can make. Even though the formula for H_1 has a negative sign, H_1 is always nonnegative, because each proportion p_i is between 0 and 1, so that $\log_2 p_i$ is negative or zero.

Next, H_1 is easy to compute in a case of equiprobability. Suppose that there are n equiprobable outcomes. Then for $i = 1, \ldots, n$ we have $p_i = \frac{1}{n}$, so

$$
H_1 = -\sum_{i=1}^{n} \tfrac{1}{n} \log_2 \tfrac{1}{n} = -n\left(\tfrac{1}{n}\log_2 \tfrac{1}{n}\right) = -\log_2 \tfrac{1}{n} = \log_2 n. \tag{1}
$$

We saw an instance of this in the example of the June garden, for which $H_1 = \log_2 4$.

3.4   Computing Base-2 Logarithms

Measures in information theory conventionally use base-2 logarithms; and because most of the data collected in this fashion have used base-2 logarithms, it is probably wise to continue to do so. Mathematically, any base would work. Base-2 logarithms were chosen early on because the original application of information theory was to problems in communication engineering which dealt with binary data (bits, each 0 or 1).

Most calculators allow you to compute the logarithm to base 10 or the natural logarithm (which uses a base of $e = 2.71828\ldots$) of a positive number. Computing the logarithm to base 2 of a positive number is not difficult with a calculator, but it cannot be done with a single keystroke. We need to use the following standard identity:

$$
\log_a p = \frac{\log_b p}{\log_b a}. \tag{2}
$$

Letting $a = 2$ and $b = 10$, we have

$$
\log_2 p = \frac{\log_{10} p}{\log_{10} 2}.
$$

For example, with your calculator you can easily check that

$$\log_2 3 = \frac{\log_{10} 3}{\log_{10} 2} \approx 1.585 \, .$$

If you prefer to use natural logarithms, let $b = e$ in **(2)**, so that

$$\log_2 b = \frac{\ln b}{\ln 2}.$$

Again with a calculator, you can check that

$$\log_2 3 = \frac{\ln 3}{\ln 2} \approx 1.585 \, .$$

Exercise

1. Use H_1 to calculate the difference in the levels of diversity represented by a garden with 9 types of equally abundant flowers versus a garden with only 5 types of flowers in equal abundance.

3.5   The Divergence from Equiprobability

Definition 2. *In an experiment with n categories, $H_{1\max}(n)$ is the maximum possible value of H_1.*

(When the number of categories is understood from the context, this quantity will be denoted by $H_{1\max}$.)

In Section 4 we will show that such a maximum always exists and in fact occurs when all n categories are equiprobable, so that H_1 satisfies the second condition of a diversity measure. Showing that H_1 satisfies the third condition of a diversity measure is left as an exercise.

The maximum value for H_1 provides a way to measure how far from equiprobability a particular set of proportions is.

Definition 3. *The divergence from equiprobability is*

$$D_1 = H_{1\max} - H_1 = \log_2 n - H_1,$$

where n is the number of categories in the system.

A low D_1 value means H_1 is close to $H_{1\max}$, that is, the system is nearly in a state of equiprobability; there is a high degree of diversity present. Conversely, a high D_1 value means that H_1 is small relative to $H_{1\max}$, that is, the system has diverged substantially from equiprobability and is not very diverse.

For example, in the August garden, only three of the four flower types were present (**Table 3**). The H_1 value for the system was calculated to be 1.5 . Thus the divergence from equiprobability in this case is

$$D_1 = H_{1\max} - H_1 = (\log_2 4) - 1.5 = 2.0 - 1.5 = 0.5 \, .$$

This is a substantial divergence, since it represents 25% of $H_{1\max}$.

Exercises

2. Use **(1)** to show that H_1 satisfies the third condition in Section 2 for a diversity measure by comparing H_1 in two different equiprobability situations: one with m outcomes and the other with n.

3. **Table 5** shows the distribution of MN blood groups in a population of Bedouins in the Syrian desert. Calculate the diversity, H_1, and the divergence, D_1, for this data set.

Table 5.

Distribution of Blood Groups in Bedouins
(adapted from [Boyd 1939, 234]).

Blood Group	MM	MN	NN
Proportion	0.57	0.37	0.06

4. **Table 6** shows the distribution of the ABO blood groups in four different populations.

Table 6.

Distribution of ABO Blood Groups
(adapted from [Boyd 1950, 223-225]).

Population	A	B	AB	O
Germans	0.425	0.145	0.065	0.365
Basques	0.417	0.011	0.000	0.572
Navajos	0.225	0.000	0.000	0.775
Chinese	0.251	0.342	0.100	0.307

   a) Using H_1 calculations, determine which population has the most diverse distribution of blood groups, and which population has the least.

   b) Determine the divergence D_1 for each of the populations.

5. In **Exercises 3** and **4**, field data are reported as arrays of proportions of times observations were made. This makes doing H_1 calculations quite easy. However, often such data are given as an array of raw frequencies of observations, as in **Table 7**.

   Calculations of H_1 can be made directly from raw frequency data, without ever converting to proportions, if rules for logarithms are used. Let $f_1, \ldots, f_n$ represent the observed frequencies of the n events $e_1, \ldots, e_n$. Let S denote the total number of observations,

$$S = \sum_{i=1}^{n} f_i = f_1 + \cdots + f_n.$$

Table 7.

Distribution of ABO Blood Groups

(adapted from [Boyd 1950, 223-225]).

Population	O	A	B	AB	Total
Siamese	79	38	75	21	213
English	202	179	35	6	422
Blackfeet Indian	27	88	0	0	115

Then the proportion of time event e_i was observed is $p_i = f_i/S$.

a) Use this expression for p_i to show that

$$H_1 = \log_2 S - \frac{1}{S} \sum_{i=1}^{n} f_i \log_2 f_i.$$

b) **Table 7** shows the distribution of ABO blood groups in three populations. Compute the measure of diversity of blood types for each of the three populations by using the formula you derived in **(a)**.

c) Which of the three groups has the highest divergence of blood types?

4.   The Proof that $H_{1\max} = \log_2 n$

4.1   Properties of $x \log_2 x$

Let us suppose now that there are n possible categories in a particular experiment, with $n > 1$. The calculation of H_1 involves summing quantities of the general form $x \log_2 x$, with x taking on each of the values p_1 through p_n, with p_i representing the proportion of observations from the ith category. To show that $H_{1\max}$ occurs when $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$, we will use elementary calculus to describe certain properties of the function $x \log_2 x$.

We begin by formalizing our convention that $x \log_2 x = 0$ when $x = 0$. We define a new function F whose domain is $0 \leq x \leq 1$:

$$F(x) = \begin{cases} 0, & \text{if } x = 0; \\ x \log_2 x, & \text{if } 0 < x \leq 1. \end{cases}$$

To make differentiation of $F(x)$ an easier task, we will employ the fact that

$$\log_2 x = \frac{\ln x}{\ln 2}.$$

If we agree to let $k = 1/\ln 2$, then we can write:

$$F(x) = \begin{cases} 0, & \text{if } x = 0; \\ kx \ln x, & \text{if } 0 < x \le 1. \end{cases}$$

Observe that $F(x)$ is continuous for $0 < x \le 1$, because both x and $\ln x$ are continuous. To see what happens at 0, we argue as follows:

$$\lim_{x \to 0^+} F(x) = \lim_{x \to 0^+} kx \ln x = \lim_{x \to 0^+} \frac{k \ln x}{1/x}.$$

Now employ l'Hôpital's Rule:

$$\lim_{x \to 0^+} \frac{k \ln x}{1/x} = \lim_{x \to 0^+} \frac{k/x}{-1/x^2} = \lim_{x \to 0^+} -kx = 0 = F(0).$$

That is, we have shown that $\lim_{x \to 0^+} F(x) = F(0)$, so F is continuous on the closed interval $[0, 1]$.

Finally, observe that for $0 < x < 1$, the first two derivatives of F are:

$$F'(x) = k + k \ln x, \tag{3}$$
$$F''(x) = k/x. \tag{4}$$

In particular, notice that since k is positive, $F''(x) > 0$. Consequently, $F'(x)$ is increasing and $F(x)$ is convex (concave up).

Exercise

 6. Graph the function F and mark any extreme points.

4.2 Using the Mean Value Theorem

The convexity of $F(x)$ turns out to be the crucial fact in determining $H_{1\max}$. To see this, recall the following result from differential calculus.

Mean Value Theorem. *If a function f is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), then there exists at least one number c in (a, b) such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Geometrically, the Mean Value Theorem means that the slope of the secant line over the entire interval is the same as the slope of the tangent line at some intermediate point in the interval (**Figure 1**).
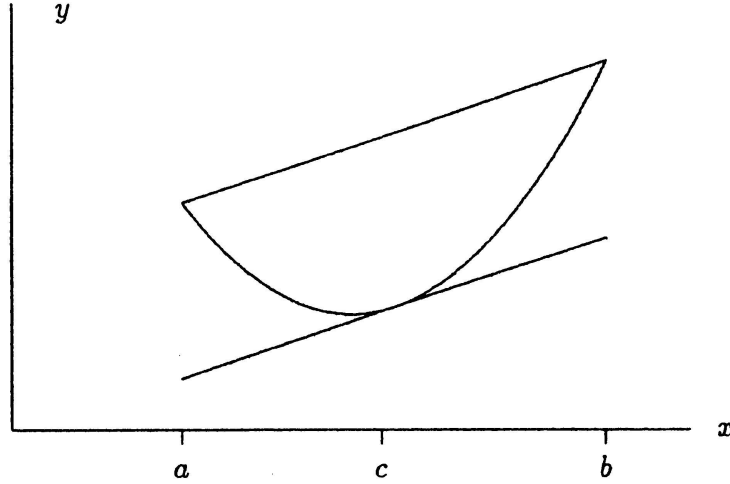
Figure 1. The Mean Value Theorem. The slope of the secant over $[a, b]$ is the same as the slope of the tangent at some intermediate point c in the interval.

Corollary 1. *Assume f is a continuous function on the closed interval $[a, b]$, differentiable on (a, b), and that f' is an increasing function. For $a \leq r < t < v \leq b$, we have*

$$\frac{f(t) - f(r)}{t - r} \leq f'(t) \leq \frac{f(v) - f(t)}{v - t}. \tag{5}$$

Proof: The Mean Value Theorem applies to f on both of the intervals $[r, t]$ and $[t, v]$. Consequently, there exist points s and u in (r, t) and (t, v), respectively, such that

$$f'(s) = \frac{f(t) - f(r)}{t - r} \qquad \text{and} \qquad f'(u) = \frac{f(v) - f(t)}{v - t}.$$

Since $s < t < u$ and f' is by hypothesis an increasing function, we have

$$f'(s) \leq f'(t) \leq f'(u)$$

and the result follows.

The geometric meaning of **Corollary 1** (see **Figure 2**) is that if f is a differentiable convex function, then the slopes of two successive secants to the curve are increasing, and the slope of the tangent at their intersection is intermediate to the slopes.

Corollary 2. *Assume f is continuous on the closed interval $[a, b]$, differentiable on (a, b), and that f' is an increasing function. Let t be any point in (a, b). Then for any point p in $[a, b]$,*

$$f(p) \geq f(t) + f'(t)(p - t). \tag{6}$$
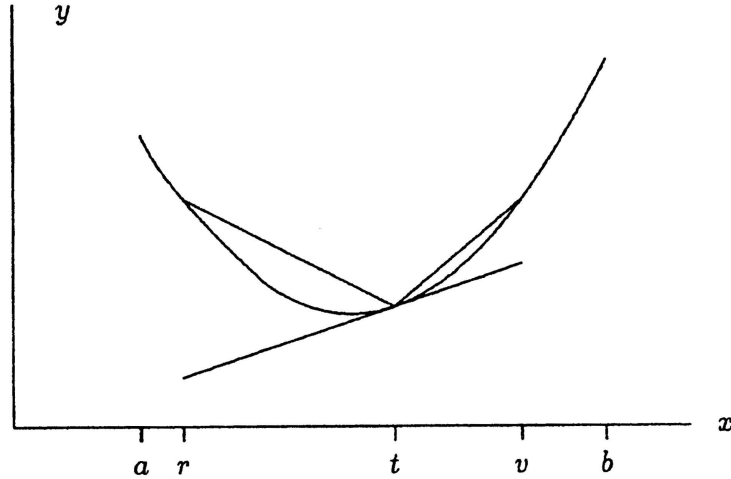
Figure 2. $f'(t)$ is intermediate to the slope of successive secants.

Proof: There are three possibilities: $p = t$, $p < t$, or $p > t$. In the case $p = t$, **(6)** reduces to the trivial assertion that $f(t) \geq f(t)$.

Now assume that $p < t$. Using p for r in the first half of **(5)**,

$$\frac{f(t) - f(p)}{t - p} \leq f'(t).$$

Consequently we have

$$f(t) - f(p) \leq f'(t)(t - p),$$

so

$$-f(p) \leq -f(t) + f'(t)(t - p).$$

Since $t - p = -(p - t)$, by multiplying this last inequality by -1, we obtain the desired result:

$$f(p) \geq f(t) + f'(t)(p - t).$$

The case $p > t$ is similar and is left as an exercise.

Corollary 2 has a simple geometric interpretation. Let f be a differentiable convex curve. Recall that the tangent line to f at t is given by

$$y - f(t) = f'(t)(x - t).$$

In particular, when $x = p$, we find

$$y = f(t) + f'(t)(p - t).$$

But **Corollary 2** says that for any point p,

$$f(p) \geq f(t) + f'(t)(p - t).$$

That is, the graph of a differentiable convex function lies above any of its tangent lines, as can be seen in either **Figure 1** or **2**.

Exercise

7. Do the case $p > t$ in the proof of **Corollary 2**.

4.3   Convexity and $H_{1\max}$

Now we can complete the proof that $H_{1\max} = \log_2 n$ for systems with n possible states. It follows from **(3)** and **(4)** in Section 4.1 that the function

$$F(x) = \begin{cases} 0, & \text{if } x = 0; \\ kx \ln x, & \text{if } 0 < x \le 1 \end{cases}$$

satisfies the conditions of **Corollary 2** on the closed interval $[0, 1]$.   Let $t = \frac{1}{n}$; then for any proportion p_i, **Corollary 2** implies

$$F(p_i) \ge F(\tfrac{1}{n}) + F'(\tfrac{1}{n}) \left(p_i - \tfrac{1}{n}\right).$$

Summing this inequality over all n probabilities for the system, we obtain

$$\sum_{i=1}^{n} p_i \log_2 p_i = \sum_{i=1}^{n} F(p_i)$$

$$\ge \sum_{i=1}^{n} \left(F(\tfrac{1}{n}) + F'(\tfrac{1}{n})(p_i - \tfrac{1}{n})\right)$$

$$= \sum_{i=1}^{n} F(\tfrac{1}{n}) + \sum_{i=1}^{n} F'(\tfrac{1}{n})(p_i - \tfrac{1}{n})$$

$$= nF(\tfrac{1}{n}) + F'(\tfrac{1}{n}) \left(\sum_{i=1}^{n} p_i - \sum_{i=1}^{n} \tfrac{1}{n}\right)$$

$$= nF(\tfrac{1}{n}) + F'(\tfrac{1}{n})(1 - 1)$$

$$= nF(\tfrac{1}{n}).$$

That is,

$$\sum_{i=1}^{n} p_i \log_2 p_i \ge nF(\tfrac{1}{n}),$$

so

$$-\sum_{i=1}^{n} p_i \log_2 p_i \le -nF(\tfrac{1}{n}),$$

and hence

$$H_1 \le -n \left(\tfrac{1}{n} \log_2 \tfrac{1}{n}\right) = \log_2 n,$$

which fact completes the proof and shows that H_1 satisfies **Condition 2** of a diversity measure in Section 2.

5.   Information Theory Applications

5.1   Three Applications to Ecological Diversity

The most common biological application of information theory is quantification of ecological diversity. When an ecosystem possesses numerous plant and animal species, with many of them present in relatively high numbers, it will have a high H_1 value and a low D_1. An ecosystem with fewer types of organisms present, or with only a few common plant or animal species, will concomitantly have lower H_1 and higher D_1 values.

We generally think of healthy biological communities in favorable habitats as being highly diverse. Decreased biological diversity may be due to environmental conditions (desert versus a temperate zone forest) or to stresses on a biological community (acid rain or pesticides), which eliminate susceptible species from the natural mix. Information theory allows us to quantify both stress-induced and natural differences between ecosystems.

Even natural changes in the the diversity of an ecosystem can be quite dramatic. Twice in the last twenty years, there have been major infestations of the crown of thorns starfish, *Acanthaster planci*, on the Great Barrier Reef. The crown of thorns attacks certain species of corals that build and maintain the reef. In some areas as much as 98% of the coral is dead, though it is not clear whether the crown of thorns is responsible for all of this destruction. Initially marine ecologists in Australia were alarmed by these infestations. However, opinion is now beginning to shift.

> Not only do some experts think the coral can recover from the crown of thorns. Some marine scientists also think it might be good for the reef to go through such growing pains. Their argument is essentially that the crown of thorns, like a renovator, may make a terrible mess on the way toward home improvements in the longer term. In particular, the crown of thorns cuts down the dominant species of coral and thus makes room for other species, currently crowded out, the upshot being a more diverse group of corals. Biologists have known for a long time that the starfish in temperate waters, by eating common things like mussels, clear out space for other species, and the same kind of benefits, they argue, may accrue from the "destruction" caused by the crown of thorns. [Ford 1988, 51]

Changes in the diversity of an ecosystem are important, as the example above indicates. Ecologists call the first-order diversity of an ecosystem the *Shannon index* [Lloyd, Zar, and Karr 1968]. An H_1 value can be computed for all of the organisms present in an environment, or for specific types of organisms such as trees or insects. Because the presence of uncommon species in nature can make information theory measures sensitive to the sizes of data sets (in general, only large sets will contain representatives of all the rare species), it is for important for comparisons of biological diversity to collect similar-sized data sets.

5.1.1   Tree Species Diversity

Table 8 contains H_1 values for mature trees found in different types of forests in Florida [Monk 1967]. Habitats of the sandhill complex type have fairly simple mature tree communities ($H_1 = 0.97$) compared to an area of sand pine scrub ($H_1 = 1.55$). The most complex mature tree community is the climax southern mixed hardwoods ($H_1 = 2.56$), into which the other forest types listed change very slowly by a process known as *ecological succession*. Information theory measures do distinguish between these different natural communities.

Table 8

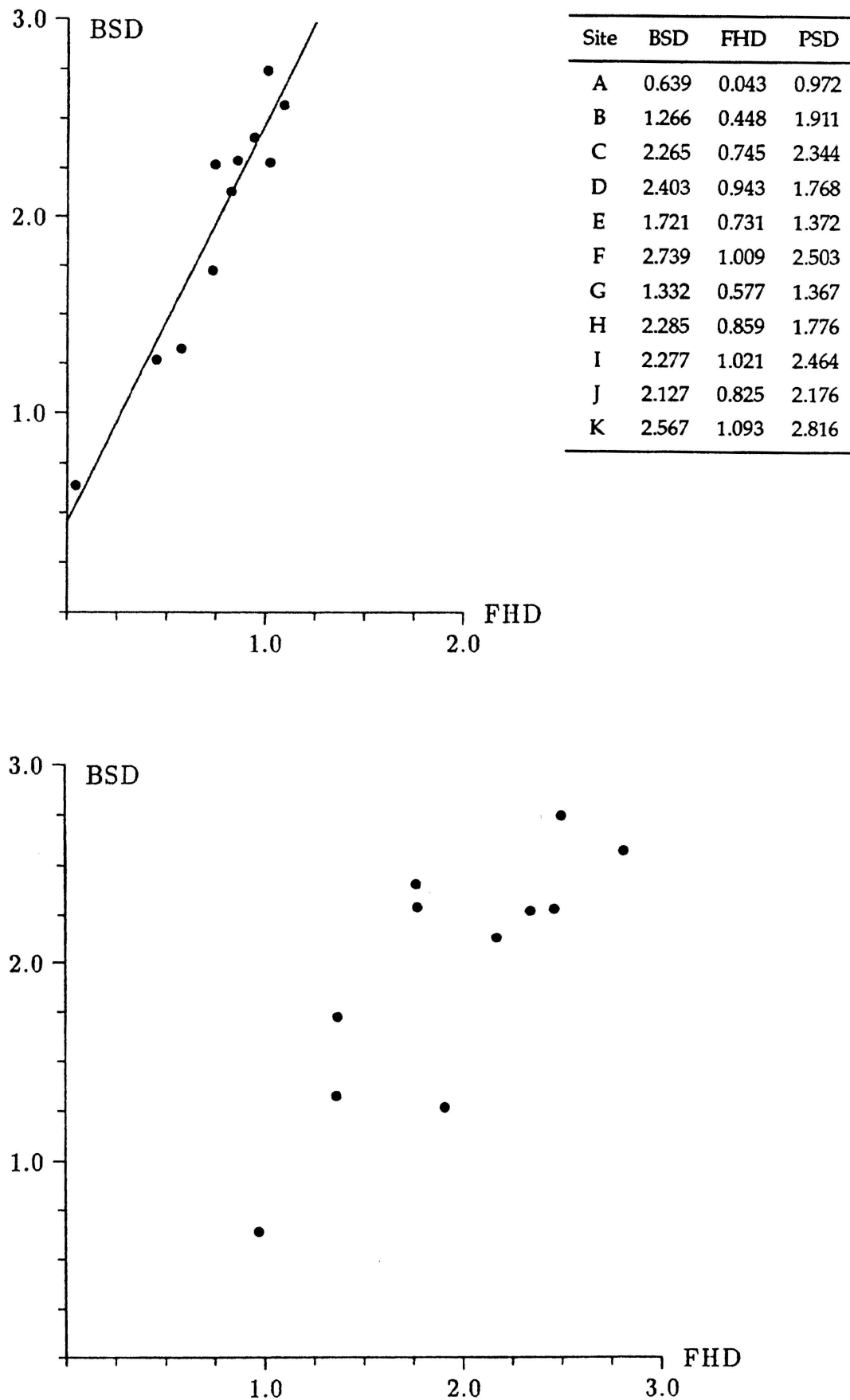H_1 values for trees in plant communities in Florida (from [Monk 1967, 175]).

Community	H_1 of Mature Trees
Sandhill Complex	0.97
Cypress Heads	1.16
Sand Pine Scrub	1.55
Mixed Hardwood Swamps	2.28
Climax Southern Mixed Hardwoods	2.56

5.1.2   Bird Species Diversity

Ornithologists have noted that more types of birds are present breeding in woodlands than in fields of similar sizes. MacArthur and MacArthur [1961] used first-order diversity measures to investigate the relationship between bird diversity and vegetation. They measured H_1 values for the diversities of bird species breeding at 11 deciduous woodland locations in Pennsylvania, Vermont, and Maryland.

In the same habitats they measured various aspects of the vegetation in order to look for any plant community characteristics that were strongly correlated with bird species diversities. Plant species diversities were computed by using H_1 values. Foliage height diversities, which expressed the number of layers of leaves between the ground and the sky in different woodlands, were also measured. Zones of 0 to 2 feet, 2 to 25 feet, and greater than 25 feet above the ground were used as height categories. The number of leaves above points on the ground were estimated for each height zone, and H_1 values for the foliage height diversity were then calculated. When the number of leaves above the ground in the three height zones are more nearly equal, the foliage height diversity measure increases to reflect the more complex physical environment.

The simplest model of how birds select a nesting habitat is that as either foliage height diversity or plant species diversity increases, the attractiveness of the habitat increases linearly. MacArthur and MacArthur [1961] looked for such a relationship. Bird species diversity and foliage height diversity were strongly correlated. **Figure 3** shows this correlation as a linear relationship.

Site	BSD	FHD	PSD
A	0.639	0.043	0.972
B	1.266	0.448	1.911
C	2.265	0.745	2.344
D	2.403	0.943	1.768
E	1.721	0.731	1.372
F	2.739	1.009	2.503
G	1.332	0.577	1.367
H	2.285	0.859	1.776
I	2.277	1.021	2.464
J	2.127	0.825	2.176
K	2.567	1.093	2.816

Figure 3. Bird species diversity (BSD) plotted against foliage height diversity (FHD) and against plant species diversity (PSD). Adapted from [MacArthur and MacArthur 1961, 596].

The data closely approximate the line given by the equation

$$\text{bird species diversity} = 2.01 \times \text{foliage height diversity} + 0.46 \,.$$

On the basis of this linear relationship, birds appear to be selecting nesting habitats on the basis of foliage height diversities.

Interestingly, there was a much weaker relationship between bird species diversity and plant species diversity measures. The plot of bird species diversity versus plant species diversity is considerably less linear than the plot using foliage height diversity. The physical structure of the woodlands, in terms of the leaves present in different height zones, seems to matter more to the birds than the species of plants producing those physical structures.

5.1.3   The Effect of Insecticide Application

Species diversity values can also be used to measure changes in a single habitat over time. **Figure 4** shows the effect of experimental insecticide application on H_1 values for arthropods in a treated grassland compared to an untreated control habitat [Barrett 1968]. The untreated control area (dashed line) does not display the dramatic drop in H_1 values immediately after application of the insecticide sevin (indicated by the shaded time period) that is seen in the treated area (solid line). A similar approach can be used to measure the effects of accidentally released pollutants.
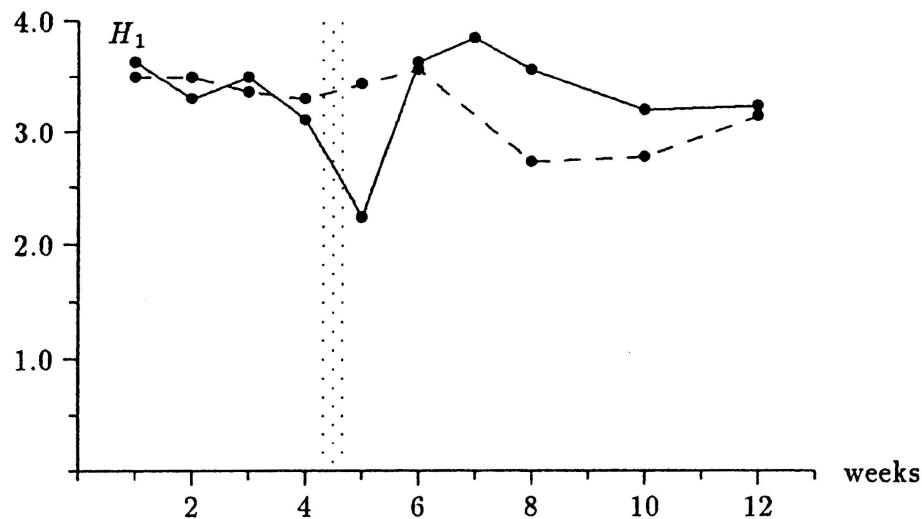
Figure 4. The effect of insecticide application (shaded region) on H_1 values of a treated habitat (solid line) and an untreated control habitat (dashed line). Modified from [Odum 1971, 150].

Measurements like the ones discussed for tree species diversity, bird species diversity, and insecticide effects, can be carried out using a number of different diversity index [Pielou 1975, 1984; Magurran 1988].

Exercises

8. Before settlement by Europeans, the Great Lakes Region of the United States was covered by a vast pine forest. Recent studies of this vanished forest area have relied on modern analysis of data collected during the General Land Office survey of Michigan carried out in the 1800s [Whitney 1986]. The data in **Table 9** show the percentage of trees reported as "bearing trees" by surveyors in two different habitats in Northern Michigan, sometime between 1836 and 1859 [Whitney 1986].

Table 9.

Percentage of species of trees reported in two forest areas in Michigan. Modified from [Whitney 1986, 1552]).

Species	Swamps	Uplands
Balsam fir	1.7	0.0
Red maple	1.6	2.6
Sugar maple	0.1	3.0
Yellow birch	0.5	0.0
White birch	2.9	0.4
Birch	2.5	0.8
Beech	0.2	17.2
White ash	0.7	0.0
Black ash	3.6	0.0
Tamarack	29.7	0.0
Spruce	10.6	0.0
Pitch pine	1.7	10.9
Norway pine	1.1	25.7
Pine	1.5	7.5
White pine	7.6	12.7
Poplar	3.8	2.2
White oak	0.0	5.7
Black oak	0.0	3.4
Cedar	25.2	0.0
Hemlock	3.2	7.3
Elm	1.0	0.1
Miscellaneous	0.8	0.5

  a) Determine the the diversity of the tree species, H_1, for each of the two areas in **Table 9**.

  b) Determine the the divergence from equiprobability, D_1, of the tree species for each of the two areas in **Table 9**.

9. Honey bees perform tasks inside the hive when they are newly-emerged workers, and switch to collecting nectar and pollen from flowers later in their lives. Numerous studies of honey bee behavior have revealed that the

tasks they perform inside the hive are quite varied and complex. The data in **Table 10** show all of the behavioral patterns recorded for three worker bees that represent a very small part of the group observed in an experiment [Kolmes 1985]. Although three bees is too small a group from which to draw any conclusions, we can still get a sense of the diversity of behavior patterns.

Table 10.

Data for honey bee behavior inside the hive. Modified from [Kolmes 1985].

Behavior	Bee 1	Bee 2	Bee 3
Walk	9	7	8
Stand	7	7	2
Groom Self	4	7	5
Inspect/Feed	1	1	2
Into Empty Cell	2	5	1
Into Honey Cell	0	1	0
Into Pollen Cell	0	1	0
Build Comb	0	3	0
Groom Other Bee	0	0	4
Get Groomed	0	0	1
Get Fed	0	0	1
DVAV	1	0	0
Attend Dance	3	0	0
Antennate	2	2	0
Chew Hive	1	0	0
Fan	1	2	0

a) Determine the the diversity of the behaviors, H_1, for each of the three bees in **Table 10**.

b) Determine the the divergence from equiprobability, D_1, of the behaviors for each of the three bees in **Table 10**.

5.2 Fire Ants

Fire ants of the species *Solenopsis saevissima* are social insects that live in underground nests containing many sterile workers and their queen. To obtain food, workers set forth from the nest and search the surrounding area. If a worker finds a food source large enough for a number of ants to harvest, a communication system based on odor trails allows additional ants to be recruited [Wilson 1962].

To produce an odor trail, a worker returning to the nest periodically drags its sting along the ground while releasing a chemical produced by Dufour's gland through the extruded sting. The chemical released is attractive to other workers

and causes them to follow the odor trail towards the food. A truly abundant food source eventually produces a situation in which many ants returning from the food to the nest are excited into producing an odor trail and the summed individual odor trails produces a virtual chemical "highway" leading to the food. The chemical secreted from Dufour's gland fades slowly over time, so that a depleted food source loses its attractiveness.

In the absence of any odor trail, a foraging worker leaving a fire ant nest might be expected to depart without bias towards any particular direction, that is, in any one of the $360°$ of directions that surround the nest. The diversity of directions that a departing group of ants might be expected to display in this uninformed initial circumstance would be

$$H_i = \log_2 360$$

in which every $1°$ of direction is taken as a potential direction category.

If there is an odor trail to a food source, then the departing ants might be expected to depart from their nest in a smaller angular range of directions. If the smaller diversity of this range of directions is symbolized by H_s, then the transmission of information by the odor trail, denoted H_t, must equal the difference between the diversity of the array of departure directions displayed by the informed ants and that displayed by the uninformed groups of ants. That is,

$$H_t = H_i - H_s.$$

Using a small drop of sugar solution placed on an index card as a food source, Wilson [1962] measured the direction indicated by the odor trail produced by a single fire ant and its influence upon the directions in which recruited foragers travelled from their nest. After this procedure was carried out a number of times to obtain replicate data sets, an estimate of the information transmitted by fire ant odor trails could be made.[2]

The results of the fire ant study showed a considerable amount of information transmission by the odor trails (see **Table 11**). When food sources were placed between 20 mm and 100 mm from an ant nest, the range of directional information transmitted by odor trails proved to be between 3 and 5 bits.

We can interpret one bit of information in this context in the following way. If a foraging ant could only inform another worker that a food source was either to the north or the south of the nest, there would be only two directional choices, so $H_{1\max} = H_i = \log_2 2 = 1$. If the communication that took place was perfect and the second worker always went in the correct direction, then $H_s = 1 \log_2 1 = 0$. In this simple situation,

$$H_t = H_i - H_s = 1.$$

---

[2]H_s can be measured either by observing the distribution of directions by which ants depart from their nest and counting the number of ants in each degree-category, or by considering the data to be a normally distributed one-dimensional Gaussian distribution and applying the formula $H_s = \log_2 \sqrt{2\pi e}\, \sigma$, with σ the standard deviation. See Haldane and Spurway [1954] or Wilson [1962] for more details concerning the latter approach.

Table 11.

The amount of directional information transmitted to single workers by a single fire ant odor trail (adapted from [Wilson 1962, 152]).

Target range (mm)	H_t
20	2.81
50	4.11
100	5.10

Now assume the foraging ant could perfectly transmit the information as to whether the food source was to the north, east, south , or west, then $H_i = \log_2 4 = 2$ bits since there are now four directional categories. In the same manner, we can interpret the 3 to 5 bits of information conveyed by the foraging ants in this experiment. The 3 to 5 bits of information transmitted is equivalent to every departing forager being told what direction to walk and being equipped with a tiny compass upon which are marked between $2^3 = 8$ and $2^5 = 32$ directional points. Departing foragers given such equipment, and able to use it well, would be able to depart in a given direction as accurately as departing ants using an odor trail rather than a tiny compass as their guide [Wilson 1962, 154].

The main reason for converting the actual field data regarding trail communication into abstract "bits of information" is that doing so allows us to compare the very different communication systems employed by a wide range of insects and animals, e.g., the information conveyed by a honey bee's waggle dance vs. that of a fire ant's odor trail.

A foraging honey bee that has located a rich source of nectar and pollen will perform a waggle dance upon its return to its hive. The dance allows the other bees to find the food source. The waggle dance looks rather like a "figure-8" performed on a vertical surface inside the hive. Its interpretation involves both the orientation of the dance and its pace. The direction to a food source is encoded by the angle that the figure-8 deviates from the vertical. The distance to the food source is encoded by the number of turns per minute that a dancing bee performs. Haldane and Spurway [1954] analyzed the information transmitted by dancing bees to other foragers. Honey bees transmit between 2.5 and 4.0 bits of directional information in their waggle dance.

Both honey bees and fire ants are quite good at directing nestmates to food sources, although the methods they use to communicate with each other differ dramatically.

5.3   Information Theory and Fish Courtship

Information theory was used by J. R. Baylis [1976] to quantify communication during the courtship of cichlid fishes. These fishes form pair bonds between male and female. After the female releases her eggs for fertilization

by the male, both parents take part in an extended period of parental care. Both the eggs and the young fry that hatch from them are guarded by the parents; and especially when the young have developed into a motile "swarm" of tiny fishes, protecting them from potential predators requires considerable efforts by both parents. Because of this vital joint responsibility, cichlid fishes have an extended period of courtship communication that allows each fish to evaluate in some fashion how good a mate and fellow-guardian its potential partner would be.

Baylis [1976] categorized the behavior patterns of the fish into 10 event categories corresponding to the various body postures and movements that make up the repertoire of courtship signals. Both males and females can carry out any of the 10 categories.

All of the fish used were young animals who were sexually mature but who had never previously spawned, so there were no previously learned aspects of social signalling to influence the behavior of the fish. The male and female fish were then placed together in an aquarium. Baylis periodically watched the fish, from initial introduction until their fry were free swimming.

The data form a preceding act/following act matrix, with 20 rows and 20 columns, with 10 columns for each behavior pattern for the male fish and 10 for the female fish. The 20 rows were labelled similarly. The data in the cells of the matrix are the frequencies with which given behavior patterns (e.g., "male performs behavior pattern number 2") were immediately followed by other behavior patterns (e.g., "female performs behavior pattern number 4"). Since communication consists of a signal and its response, this arrangement records signal–response pairs. By having the sexes on both rows and columns, it was possible to record an animal receiving a reply to its signal (that is, a male responded to by a female, or vice versa) and also an animal "repeating itself" (that is, a male or female sending 2 signals in a row, without an intervening reply from the other fish). If you consider for a moment how frequently humans must repeat questions before being answered, it will be obvious why the matrix was arranged to allow recording of repetition.

Each matrix is for a particular time in the sequence of courtship behavior. The entire set of observations is a set of data matrices, each labelled with the time of recording (e.g., "1 hour after the fish were placed together in the observation aquarium").

Table 12. A transition array of preceding (rows) and following (columns) behaviors observed during observation period 1 for three pairs of Cichlasoma citrinellum. Adapted from Baylis [1976, 125].

Preceding Act / Following Act

Behavior	A	B	C	D	E	F	G	H	I	J	a	b	c	d	e	f	g	h	i	j
						Male										Female				
Male																				
A. Bite	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
B. Frontal	0	0	6	0	0	2	0	0	0	1	0	5	9	0	1	1	0	5	1	1
C. Lateral	0	0	19	4	76	76	0	1	10	86	0	13	53	11	3	65	1	1	61	122
D. Dig	0	0	2	1	1	0	0	0	0	2	0	1	1	0	0	6	0	0	1	5
E. Tailbeat	1	0	9	0	87	12	0	0	0	21	0	5	6	3	1	7	0	2	16	30
F. Quiver	0	0	22	2	3	184	0	0	26	24	0	1	0	8	0	95	0	0	54	36
G. Skim	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H. Yield	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
I. Nip Off	0	0	4	3	1	19	0	0	35	9	0	0	0	0	0	15	0	0	5	27
J. Approach	0	17	73	1	6	9	0	0	5	12	0	4	13	10	0	79	0	8	61	63
Female																				
a. Bite	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
b. Frontal	0	7	116	0	2	0	0	0	1	0	0	0	3	0	0	1	0	0	0	3
c. Lateral	0	0	38	0	5	2	0	0	1	7	0	2	5	1	2	8	0	1	6	26
d. Dig	0	0	8	2	1	7	0	0	1	12	0	0	0	12	0	27	0	0	1	14
e. Tailbeat	0	0	3	0	3	0	0	0	0	1	0	0	1	0	6	1	0	0	0	0
f. Quiver	0	0	49	5	3	80	0	0	25	67	0	0	2	27	1	287	1	1	66	65
g. Skim	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	2	0
h. Yield	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	5	0	0	7
i. Nip Off	0	2	39	0	10	28	0	0	5	87	1	0	1	11	1	56	0	1	317	38
j. Approach	0	6	214	2	2	36	0	0	9	25	102	7	7	1	0	26	0	0	6	16

An inspection of the matrix in **Table 12** tells us a bit about what is going on. For example, males often followed one "quiver" display with another (row F, column F). Roughly half as often, a female responds with a "quiver" to a male's "quiver" display (row F, column FF). But this approach to analyzing the matrix is a tedious one; what can information theory tell us here?

We can use row and column totals to compute first-order information theory measures. As courtship proceeds, we would expect to see the divergence from equiprobability (D_1) increase for a pair of fish. The signals sent by each sex may begin with a high first-order uncertainty (H_1). However, unless the fish begin to abandon signals unnecessary to spawning and to concentrate on the final spawning signals (therefore reducing H_1 and increasing D_1) they will never produce young together.

In a human analogy, unless their conversation abandons topics such as the weather and television, two people are never going to learn enough about each other to decide whether or not a mutual attraction and set of shared values exists. The divergence from equiprobability of their conversation must increase, that is, they must move beyond initial polite phrases ("Nice day, isn't it?") if anything more involved than "small talk" is to occur eventually.

The first-order measures can be computed by summing the preceding acts (rows) or the following acts (columns) and treating the total values as two one-dimensional 20-category matrices. The quantities H_1, H_{1max}, and D_1 are then computed as in any other example.

Baylis [1976] found for pairs of cichlid fish that the first-order measures (increasing D_1 values) reflected the process of courtship communication. Notice in **Figure 5** the way the first-order measures are clearly structured around the moment of spawning, with D_1 increasing just before spawning. The fish do begin to concentrate on the use of specific signals immediately before the female releases her eggs and the male releases sperm to fertilize them. Animals with external fertilization of this sort must be very well behaviorally synchronized if the male and female are both to release their gametes at the same moment.

The approach to animal communication just demonstrated does not, in the final analysis, allow us to achieve the ultimate goal of "talking to the animals." It does, however, provide one of the best means devised to date of quantifying animal communication. It is a bit amusing, however, that information theory, with its origins in devising communication systems, allows us to "listen in" (however imperfectly) to "conversations" between members of other species.

Exercises

10. The data in **Table 13** show how responsive different female white-crowned sparrows (*Zonotrichia leucophrys*) were to the recorded courtship songs of male sparrows of two species raised in the laboratory. Responses to their own species are in the top row and responses to a related species (song sparrow) are in the bottom row. Animal communication must follow predictable patterns for it to be effective, so we might expect a more consistent response by female sparrows to appropriate (same species) male songs
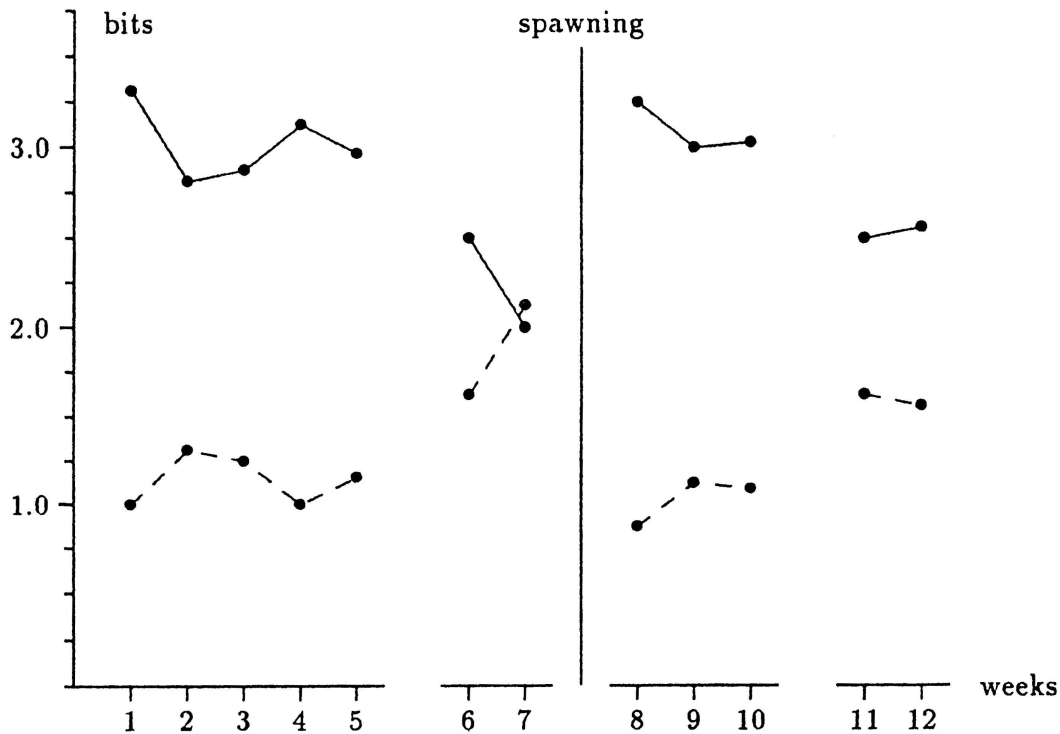
Figure 5. First-order diversity, H_1, (solid line) and divergence from equiprobability, D_1, (dashed line) of courtship during 12 observation periods in three pairs of *Cichlasoma citrinellum*. Adapted from [Baylis 1976, 134].

than to inappropriate (different species) male courtship songs. Information theory measures are one way of addressing how consistently a group of experimental animals responds to different stimuli.

a) Determine which of the laboratory-raised male sparrows elicited a more diverse response from the 13 female white-crowned sparrows.

b) What is the divergence for each data set? From just looking at the data, do you expect this divergence to be high or low (as a proportion of $H_{1\max}$)?

Table 13.

Responses by 13 female white-crowned sparrows to male white-crowned sparrows (W-CS) and male song sparrows (SS). Adapted from [Spitler-Nabors and Baker 1987, 383].

Female	a	b	c	d	e	f	g	h	i	j	k	l	m
W-CS	1	4	10	1	4	20	1	1	1	7	14	1	3
SS	2	0	10	0	0	14	27	0	0	0	10	0	4

6.   Sample Projects

We have discussed a number of applications of information theory measures. It is your turn to design an independent project in which you will collect data and use information theory measures in your analysis. You may choose to make a local insect collection, to examine the types of insect pollinators that visit various flowers, to measure bird diversity or plant diversity, to examine conversational topics or the types of clothing worn by your fellow students. We encourage you to develop an entirely novel topic. Should the size of your proposed project make it reasonable, a team effort might be appropriate.

You may want to compare two systems, as in Section 5.1.1. Or you might design a project which compares two or more different diversity indices to each other, as in Section 5.1.2. You may even want to compare one system to itself over time, as in Sections 5.1.3 and 5.3. Whatever you select, information theory measures are widely applicable; and with a little imagination you should be able to come up with an interesting topic. Birds, bees, flowers, friends, food—anything that exists in a naturally diverse array is a good place to look for a topic. Below, we outline the procedures for a few sample projects.

Here are a few reminders before you begin. First, the categories you use must not overlap and together must include all possible events. Second, you should have a well-defined sampling procedure (devised beforehand) that you follow to collect your data, to prevent unconscious biases from entering your data set. Third, it may be simpler to do your diversity calculations with raw frequencies (as in **Exercise 5**) rather than converting your data to proportions.

6.1   Diversity of Flowers

Select two or more different areas of open field. Try to select areas that look different in an overall way to you, such as a field of clover and Queen Anne's lace, and a small meadow in the middle of the woods filled with spring ephemeral flowers. If you select areas that differ in a visual and intuitive way, you can go on to see whether or not they differ in a quantifiable fashion. In each area carry out the following procedure.

1. Start on one edge of of the field and begin to walk straight towards some obvious landmark.

2. Every two paces, stop and drop in front of you a 30 cm × 30 cm square made of pipe cleaners.

3. Examine all of the flowers inside the square. Assign each one to a category, and keep track of how many flowers of each type you see. Carrying a notebook to which you tape a single representative of each of the flower categories can help you keep them straight. If you are artistically inclined, you may wish to make sketches.

4. Note that you do not need to know the names of the flowers. So long as your categories are unambiguous and include all of the types of flowers you run across, they will allow you to make correct diversity calculations.

5. After sampling each of the fields, add the numbers of flowers seen in each area. If the numbers are approximately equal, no further sampling is required. Return to any underrepresented field and walk across it on other paths until all of your fields have roughly equal sample sizes. (Each area should provide at least 100 flowers).

6. Calculate H_1, $H_{1\mathrm{max}}$, and D_1 for each field and compare the results.

6.2   Leaf Shape Diversity

Forests look very different from one another depending on where on earth they are. The Eastern deciduous forest is full of the broad leaves of maple and oak, the temperate rain forest of the Pacific Northwest is all evergreen trees and ferns, and the tropics are filled with an amazing variety of plants. You can calculate leaf shape diversity indices for habitats near you by carrying out the following procedure.

1. Go to a habitat that you select for the initial measurements.

2. Walk into the habitat two paces, close your eyes, and look downwards (or in a direction appropriate to the habitat). Open your eyes; the first plant you focus upon will be the one you use next in the data collection.

3. Find a leaf on that first plant, attach it to a blank sheet of paper and mark a "|"next to it. Take two paces and repeat step 2 above.

4. Examine the new leaf to see if it matches the other one. If it does, add another mark next to the previously attached leaf. If not, attach the new leaf and mark a "|"next to it.

5. Repeat steps 2–5 above until you have scored 100 leaves per habitat.

6. Calculate H_1, $H_{1\mathrm{max}}$, and D_1 for each area selected and compare results. What might account for any differences in these values?

6.3   Student Migrations

Animals migrate with the seasons, as the geese flying south every autumn remind those of us in central New York. Other areas are the scenes of other migrations, like the spectacular flights of monarch butterflies on the west coast as they travel to Mexico. Students also migrate on an annual cycle, from homes in various places to their colleges. You can determine how diverse the set of geographical origins of your institution's student body is by carrying out the following procedure.

1. Go to a student parking lot on your campus. If there are no specifically student-assigned lots, select one that is near the dormitories and as far as possible from any other academic buildings.

2. Walk along a row of cars, scoring the license plates by state or province of origin. Collect scores for 100 or more cars.

3. Carry out a procedure similar to step 2 above, but in the parking lot for a grocery store or mall away from campus. It is usually a good assumption that members of the local population will frequent such a store.

4. Calculate H_1, $H_{1\mathrm{max}}$, and D_1 values for your two data sets and compare the results. What does it tell you about student migrations, and why did you need to collect a second non-student sample in order to interpret your data sets well?

5. Are there any portions of the student population that might be under or over represented by the method used to collect data in step 2? Often, your college catalog, admissions office, or registrar can supply accurate data on the geographical origins of the student population. If you can obtain such data, use it to compute H_1, $H_{1\mathrm{max}}$, and D_1 for the student body and compare the results to your earlier calculations.

7.   Solutions to the Exercises

1. $\log_2 9 - \log_2 5 \approx 0.848$.

2. Assume $m > n$, then $m/n > 1$, so $\log_2(m/n) > 0$. Now, if both systems are in states of equiprobability, then

$$H_1(m) - H_1(n) = \log_2 m - \log_2 n = \log_2(m/n) > 0.$$

3. $H_1 \approx 1.237$ bits; $D_1 \approx 0.348$ bits.

4. a) Most diverse: Chinese, $H_1 \approx 1.885$ bits; least diverse: Navajos, $H_1 \approx 0.769$ bits.

   b) Germans: $D_1 \approx 0.284$ bits; Basques: $D_1 \approx 0.941$ bits; Navajos: $D_1 \approx 1.231$ bits; Chinese: $D_1 \approx 0.115$ bits.

5. a) Siamese: $H_1 \approx 1.834$ bits; English: $H_1 \approx 1.419$ bits; Blackfeet Indians: $H_1 \approx 0.786$ bits.

   b) The Blackfeet Indians have the highest divergence, $D_1 \approx 1.214$ bits.

6.

$(e^{-1}, -ke^{-1})$
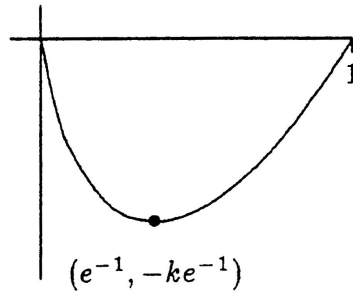
8. a) Swamps: $H_1 \approx 3.14$ bits; uplands: $H_1 \approx 3.17$ bits.

   b) Swamps: $D_1 \approx 1.32$ bits; uplands: $D_1 \approx 1.29$ bits.

9. a) Bee 1: $H_1 \approx 2.86$ bits; bee 2: $H_1 \approx 2.97$ bits; bee 3: $H_1 \approx 2.60$.

   b) Bee 1: $D_1 \approx 1.14$ bits; bee 2: $H_1 \approx 1.03$ bits; bee 3: $D_1 \approx 1.40$.

10. a) White-crowned sparrows: $H_1 \approx 2.95$ bits; song sparrows: $H_1 \approx 2.21$ bits;

    b) White-crowned sparrows: $D_1 \approx 0.75$ bits; song sparrows: $D_1 \approx 1.49$ bits.

References

Barrett, G. W. 1968. The effects of an acute insecticide stress on a semi-enclosed grassland ecosystem. *Ecology* 49: 1019–1035.

Baylis, J. R. 1976. A quantitative study of long-term courtship: II. A comparative study of the dynamics of courtship in two new world cichlid fishes. *Behaviour* 59: 118–161.

Boyd, W. C. 1939. Blood Groups. *Tabulae Biologicae* 17: 234.

————. 1950. *Genetics and the Races of Man.* Lexington, MA: D. C. Heath.

Ford, D. 1988. A reporter at large: crown of thorns. *New Yorker* (25 July 1988): 34–63.

Haldane, J .B. S. and H. Spurway. 1954. A statistical analysis of communication in *Apis mellifera* and a comparison with communication in other animals. *Insectes Sociaux* 1: 247–283.

Khinchin, A. I. 1957. *Mathematical Foundations of Information Theory.* New York: Dover.

Kolmes, S. A. 1985. An information-theory analysis of task specialization among worker honey bees performing hive duties. *Animal Behaviour* 33: 181-187.

Lloyd, M., J. H. Zar, and J. R. Karr. 1968. On the calculation of information theoretical measures of diversity. *American Midland Naturalist* 79: 257–272.

MacArthur, R. H., and J. W. MacArthur. 1961. On bird species diversity. *Ecology* 42: 594–598.

Magurran, A. E. 1988. *Ecological Diversity and its Measurement*, Princeton, NJ: Princeton University Press.

Monk, C. D. 1967. Tree species diversity in the eastern deciduous forest with particular reference to north central Florida. *American Naturalist* 101: 173–187.

Odum, E. P. 1971. *Fundamentals of Ecology*, 3rd ed. Philadelphia: W. B. Saunders.

Pielou, E. C. 1975. *Ecological Diversity*. New York: Wiley.

Pielou, E. C. 1984. *The Interpretation of Ecological Data*. New York: Wiley.

Spitler-Nabors, and Baker. 1987. Sexual display response of female white-crowned sparrows to normal, isolate, and modified conspecific songs. *Animal Behaviour* 35: 380–386.

Whitney, G. G. 1986. Relation of Michigan's presettlement pine forests to substrate and disturbance history. *Ecology* 67: 1548–1559.

Wilson, E. O. 1962. Chemical Communication among workers of the fire ant *Solenopsis saevissima* (Fr. Smith). 2. An information analysis of the odor trail. *Animal Behaviour* 10: 148–158.

About the Authors

Steven Kolmes received his B.S. in zoology from Ohio University and his M.S. and Ph.D. degrees in zoology from the University of Wisconsin at Madison. He currently holds the Rev. John Molter, C.S.C., Chair in Science at the University of Portland. He is interested in behavioral ecology at the pest–pesticide interface and efficiency theory in social insects. When this Module was written, he was chair of the Biology Dept. at Hobart and William Smith Colleges.

Kevin Mitchell received his B.A. in mathematics and philosophy from Bowdoin College and his Ph.D. in mathematics from Brown University. His main areas of interest are hyperbolic tilings, algebraic geometry, and applications of mathematics to environmental science.

Hobart and William Smith Colleges' emphasis on and support for multidisciplinary work permitted both authors to explore new areas, including team-teaching a course entitled "Mathematical Models and Biological Systems," for which this unit was developed.