Chapter 6 *Multidimensional qualitative data*

6.0 General principles

Ecologists often use variables that are neither quantitative nor ordered (Table 1.2). Variables of this type may be of physical or biological nature. Examples of qualitative physical descriptors are the colour, locality, geological substrate, or nature of surface deposits. Qualitative biological descriptors include the captured or observed species, where the different states of the nonordered descriptor are the different possible species. Likewise, the presence or absence of a species cannot, in most cases, be analysed as a quantitative variable; it must be treated as a semiquantitative or qualitative descriptor. A third group of qualitative descriptors includes the results of classifications — for example, the biological associations to which the zooplankton of various lakes belong, or the chemical groups describing soil cores. Such classifications, obtained or not by clustering (Chapter 8), define qualitative descriptors and, as such, they are amenable to numerical interpretation (see Chapter 10).

The present chapter discusses the analysis of *qualitative* descriptors; methods appropriate for bivariate and multivariate analysis are presented. Because information theory is an intuitively appealing way of introducing these methods of analysis, Section 6.1 shows how to measure the amount of information in a qualitative descriptor. This paradigm is then used in the following sections.

Contingency table

The comparison of qualitative descriptors is based on *contingency tables*. In order to compare two qualitative descriptors, the objects are first allocated to the cells of a two-way contingency table whose rows and columns respectively correspond to the two descriptors. In such a table, the number of rows is equal to the number of states of the first descriptor and the number of columns to that of the second descriptor. Any cell in the table, at the intersection of a row and a column, corresponds to one state of each descriptor. The number of objects with these two states is recorded in the cell, hence the values in contingency tables are *frequencies*. The analysis of *two-way contingency tables* is described in Section 6.2. When there are more than two descriptors, *multiway*

(or *multidimensional*) *contingency tables* are constructed as extensions of two-way tables. Their analysis is discussed in Section 6.3. Finally, Section 6.4 analyses the *correspondence* between descriptors in a contingency table.

Contingency table analysis is the qualitative equivalent of both *correlation analysis* and *analysis of variance*; in the particular case of a two-way contingency table, the analysis is the equivalent of a one-way ANOVA. It involves the computation of X^2 (chi-square) statistics or related measures, instead of correlation or F-statistics. Two types of null hypotheses (H_0) may be tested. The first one is the independence of the two descriptors, which is the usual null hypothesis in correlation analysis (H_0: the correlation coefficient $\rho = 0$ in the statistical population). The second type of hypothesis is similar to that of the analysis of variance. In a two-way contingency table, one of the descriptors (called "first descriptor" in the next sentence) corresponds to the classification criterion of the analysis of variance, and the other descriptor (called "second descriptor") corresponds to the dependent variable. The analysis compares, among the states of the first descriptor, the distribution of frequencies among the states of the second descriptors. The null hypothesis says that the frequency distributions are the same, i.e. that the observations form a homogeneous group. For example, if the groups (classification criterion) form the columns whereas the dependent variable is in the rows, H_0 states that the frequency distributions of the row frequencies are the same in all columns. These two types of hypotheses require the calculation of the same expected values and the same test statistics. The examples in the present chapter will be formulated as correlation hypotheses. In multiway tables, the hypotheses tested are often quite complex because they take into account interactions among the descriptors (Section 6.3).

Correlation hypothesis

ANOVA hypothesis

Considering species data, the names of the various species observed at a sampling site are the states of a qualitative multi-state descriptor. Section 6.5 will discuss *species diversity* as a measure of dispersion of this qualitative descriptor.

The mathematics used throughout this chapter are quite simple and require no prior knowledge other than the intuitive notion of probability. Readers interested in applications only may skip Section 6.1 and come back to it when necessary. To simplify the notation, the following conventions are followed throughout the chapter. When a single descriptor is considered, this descriptor is called **a** and its states have subscripts i going from 1 to q, as in Fig. 1.1. In two-way contingency tables, the descriptors are called **a** and **b**. The states of **a** are denoted a_i with subscripts i varying from 1 to r (number of rows), while the states of **b** are denoted b_j with subscripts j varying from 1 to c (number of columns).

6.1 Information and entropy

Chapters 1 and 2 have shown that the ecological information available about the objects under study is usually (or may be reformulated as) a set of biological and/or

environmental characteristics, which correspond to as many descriptors. Searching for groups of descriptors that behave similarly across the set of objects, or that may be used to forecast one from the other(s) (R analysis, Section 7.1), requires measuring the *amount of information* that these descriptors have in common. In the simplest case of two descriptors **a** and **b** (called $\mathbf{y}_1$ and $\mathbf{y}_2$ in previous chapters), one must assess how much *information* is provided by the distribution of the objects among the states of **a**, that could be used to forecast their distribution among the states of **b**. This approach is central to the analysis of relationships among ecological phenomena.

In 1968, Ludwig von Bertalanffy wrote, in his *General System Theory* (p. 32): "Thus, there exist models, principles, and laws that apply to generalized systems or their subclasses, irrespective of their particular kind, the nature of their component elements, and the relations or 'forces' between them". This is the case with information, which can be viewed and measured in the same manner for all systems. Some authors, including Pielou (1975), think that the concepts derived from information theory are, in ecology, a model and not a homology. Notwithstanding this opinion, the following sections will discuss how to measure information for biological descriptors in terms of information to be acquired, because such a presentation provides a better understanding of the nature of information in ecological systems.

The approach consists in measuring the amount of information contained in each descriptor and, further, the amount of information that two (or several) descriptors have in common. If, for example, two descriptors share 100% of their information, then they obviously carry the same information. Since descriptors are constructed so as to partition the objects under study into a number of states, two descriptors have 100% of their information in common when they partition a set of objects in exactly the same way, i.e. into equal and corresponding sets of states. When descriptors are qualitative, this correspondence does not need to follow any ordering of the states of the two descriptors. For ordered descriptors, the ordering of the correspondence between states is important and the techniques for analysing the information in common belong to correlation analysis (Chapters 4 and 5).

Entropy

The mathematical theory of information is based on the concept of *entropy*. Its mathematical formulation was developed by Shannon (Bell Laboratories) who proposed, in 1948, the well-known equation*:

$$H = -\sum_{i=1}^{q} p_i \log p_i \qquad \textbf{(6.1)}$$

* This equation is sometimes referred to as the Shannon-Weaver or the Shannon-Wiener equation. Norbert Wiener had developed elements of probability theory that were used by Claude E. Shannon in his 1948 paper. In 1963, Warren Weaver co-authored with Shannon a book where Shannon's 1948 article was reprinted.

Table 6.1 Contingency table (numerical example). Distribution of 120 objects on descriptors **a** and **b**.

	b_1	b_2	b_3	b_4
	30	30	30	30
$a_1 = 60$	30	10	15	5
$a_2 = 30$	0	20	0	10
$a_3 = 15$	0	0	0	15
$a_4 = 15$	0	0	15	0

where H is a measure of the uncertainty or choice associated with a frequency distribution (vector) **p**; p_i is the probability that an observation belongs to state i of the descriptor (Fig. 1.1). In practice, p_i is the proportion (or relative frequency, on a 0-1 scale) of observations in state i. Shannon recognized that his equation was similar to the equation of entropy, published in 1898 by physicist Boltzmann as a quantitative formulation of the second law of thermodynamics, which concerns the degree of disorganization in closed physical systems. He thus concluded that H corresponds to the entropy of information systems.

Negative entropy

The entropy of information theory is actually the *negative entropy* of physicists. In thermodynamics, an increase in entropy corresponds to an *increase in disorder*, which is accompanied by a *decrease of information*. Strictly speaking, information is negative entropy and it is only for convenience that it is simply called entropy. *In information theory, entropy and information are taken as synonymous*.

Information

Numerical example. In order to facilitate the understanding of the presentation up to Section 6.4, a small numerical example will be used in which 120 objects are described by two descriptors (**a** and **b**) with 4 states each. The question is to determine to what extent one descriptor can be used to forecast the other. The data in the numerical example could result from the survey of 120 sites of an estuary, or the trees observed in 120 vegetation quadrats. Descriptor **a** could be the dominant species at each sampling site, assuming there are 4 possible species, and descriptor **b**, some environmental variable with 4 states. The following discussion is valid for any type of qualitative descriptor as well as for ordered descriptors divided into classes.

Assume that the 120 observations are distributed as 60, 30, 15 and 15 among the 4 states of descriptor **a** and that there are 30 observations in each of the 4 states of descriptor **b**. The frequencies in the combined states of the descriptors (i.e. the table cells) are shown in Table 6.1.

For each descriptor, the probability of a state is estimated by the relative frequency with which the state is found in the set of observations. Thus, the probability distributions associated with descriptors **a** and **b** are:

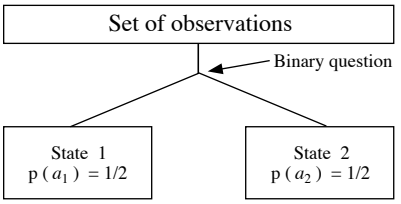
a_1:	60 p(a_1) = 1/2	b_1:	30 p(b_1) = 1/4
a_2:	30 p(a_2) = 1/4	b_2:	30 p(b_2) = 1/4
a_3:	15 p(a_3) = 1/8	b_3:	30 p(b_3) = 1/4
a_4:	15 p(a_4) = 1/8	b_4:	30 p(b_4) = 1/4
	120		120

The relative frequency of a given state is the probability of observing that state when taking an object at random.

Within the framework of information theory, the entropy of a probability distribution is measured, not in kilograms, metres per second, or other such units, but in terms of decisions. The measurement of entropy must reflect how difficult it is to find, among the objects under study, one that has a given state of the descriptor. An approximate measure of entropy is the average minimum number of binary questions that are required for assigning each object to its correct state. Hence, the *amount of information* gained by asking binary questions, and answering them after observing the objects, is equal to the *degree of disorder* or *uncertainty* initially displayed by the frequency distribution. Given that context, the terms *entropy* and *information* are used synonymously. A few numerical examples will help understand this measure.

1. When all the objects exhibit the same state for a descriptor, everything is known *a priori* about the distribution of observations among the different states of the descriptor. There is a single state in this case; hence, the number of binary questions required to assign a state to an object is zero ($H = 0$), which is the minimum value of entropy.

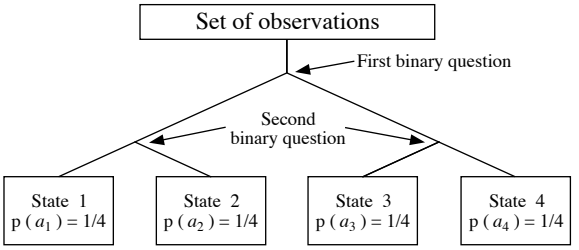
2. The simplest case of a descriptor with non-null entropy is when there are two states among which the objects are distributed equally:



Binary question

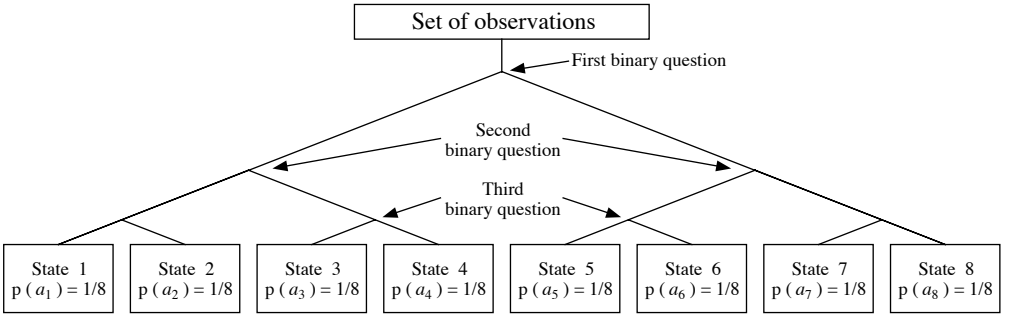
In order to assign a state to any given object, a single binary question is necessary, of the type "Does this object belong to state 1?" If it does, state 1 is assigned to the object; if it does not, the object belongs to state 2. The entropy associated with the descriptor is thus $H = 1$.

3. Applying the above approach to a descriptor with four states among which the objects are distributed equally, one gets an entropy $H = 2$ since exactly two binary questions are required to assign a state to each object:



This would be the case of descriptor **b** in the numerical example of Table 6.1.

4. For an eight-state descriptor with the objects equally distributed among the states, the binary questions are as follows:



The total entropy of the descriptor is thus:

$$[3 \text{ questions} \times 8 \text{ (1/8 of the objects)}] = 3$$

and, in general, the entropy H associated with a descriptor in which the objects are equally distributed among states is equal to the base 2 logarithm (if the questions are binary) of the number of states:

$\log_2 1 = 0$	$\log_2 8 = 3$
$\log_2 2 = 1$	$\log_2 16 = 4$
$\log_2 4 = 2$	etc.

Hence the general formula in that case is $H = \log_2(\text{number of states})$.

Measuring the entropy from the number of binary questions is strictly equal to the logarithmic measure only when the number of states is an integer power of 2, or when the number of observations in the various states is such that binary questions divide them into equal groups (see the numerical example, below). In all other cases, the number of binary questions required is slightly larger than $\log_2(\text{number of states})$,

Table 6.2 The average minimum number of binary questions required to remove the uncertainty about the position of an object in the state-vector is equal to $\log_2$(number of states) when the number of states is an integer power of 2 (in boldface) and the objects are equally distributed among the states. In all other cases, the number of binary questions is slightly larger than the entropy $H = \log_2$(number of states). For example, for a three-state descriptor with equal frequencies, the minimum number of binary questions is (2 questions × 2/3 of the objects) + (1 question × 1/3 of the objects) = 1.66666 binary questions.

Number of states	$\log_2$(number of states)	Average minimum number of binary questions
1	**0.00000**	**0.00000**
2	**1.00000**	**1.00000**
3	1.58496	1.66666
4	**2.00000**	**2.00000**
5	2.32193	2.40000
6	2.58496	2.66666
7	2.80735	2.85714
8	**3.00000**	**3.00000**
9	3.16993	3.22222
10	3.32193	3.40000
11	3.45943	3.54545
12	3.58496	3.66666
13	3.70044	3.76154
14	3.80735	3.85714
15	3.90689	3.93333
16	**4.00000**	**4.00000**

because binary questions are then a little less efficient than in the previous case (Table 6.2). Binary questions have been used in the above discussion only to provide readers with a better understanding of entropy, the true measure being the logarithmic one. One may refer to Shannon (1948), or a textbook on information theory, for a more formal discussion of the measure of entropy.

The following example illustrates the relationship between probability and information. If an ecologist states that water in the Loch Ness is fresh, this is trivial since the probability of the event is 1 (information content null). If, however, he/she announces that she/he has captured a specimen of the famous monster, this statement contains much information because of its low probability (the dynamic aspects of Loch

Ness Monster populations have been discussed by Sheldon & Kerr, 1972, Scheider & Wallis, 1973, and Rigler, 1982; see also Lehn, 1979, and Lehn & Schroeder, 1981, for a physical explanation of the Loch Ness and other aquatic monsters). Thus, information theory deals with a specific technical definition of information, which may not correspond to the intuitive concept. A nontechnical example is that a book should contain the same amount of information before and after one has read it. From the information theory point of view, however, after one has read the book once, there is no information to be gained the next time he/she reads it — unless she/he has forgotten part of it after the first reading.

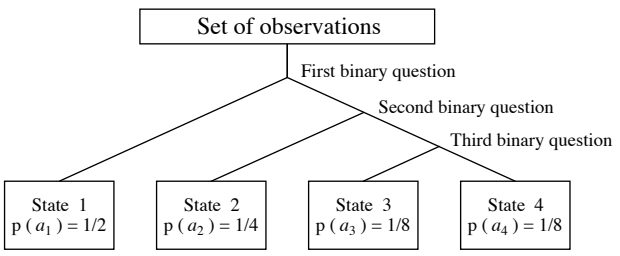
It should be clear, at this point of the discussion, that the entropy of a descriptor depends, among other characteristics, on the number of its states among which the entropy is partitioned. In the case of the above four-state descriptor, for example, 1/4 of the entropy of the descriptor is attributed to each state, i.e. $[1/4 \log_2 4]$, which is equal to $[1/4 \log_2(1/4)^{-1}]$. The total entropy of the descriptor is thus:

$$H = \sum_{4 \text{ states}} (1/4) \log_2 (1/4)^{-1} = \log_2 4 = 2$$

The same holds for the example of the eight-state descriptor. The entropy of each state is $[1/8 \log_2 8] = [1/8 \log_2(1/8)^{-1}]$, so that the total entropy of the descriptor is

$$H = \sum_{8 \text{ states}} (1/8) \log_2 (1/8)^{-1} = \log_2 8 = 3$$

5. Descriptor **a** in the numerical example (Table 6.1) illustrates the case of a descriptor for which the objects are not equally distributed among states. The probability distribution is [1/2, 1/4, 1/8, 1/8], which corresponds to the following scheme of optimal binary questions:



When the objects are not distributed evenly among the states, the amount of information one has *a priori* is higher than in the case of an even distribution, so that the information to be acquired by actual observation of the objects (i.e. the entropy) decreases. It follows that the entropy of the above descriptor should be $H < 2$, which is the maximum entropy for a four-state descriptor. Using binary questions, it is more economical to isolate half of the objects with the first question, then half of the remaining objects with the second question, and use a third question for the last two groups of 1/8 of the objects (see above). Since half of the objects require one question, 1/4 require 2, and the two groups of 1/8 require 3, the total entropy of this descriptor is:

$$H(\mathbf{a}) = (1/2 \times 1) + (1/4 \times 2) + (1/8 \times 3) + (1/8 \times 3) = 1.75$$

As in the previous examples, this is equal to:

$$H(\mathbf{a}) = 1/2\ 1\log_2 2 + 1/4\ 1\log_2 4 + 1/8\ 1\log_2 8 + 1/8\ 1\log_2 8$$

$$H(\mathbf{a}) = 1/2\ 1\log_2 (1/2)^{-1} + 1/4\ 1\log_2 (1/4)^{-1} + 1/8\ 1\log_2 (1/8)^{-1} + 1/8\ 1\log_2 (1/8)^{-1}$$

$$H(\mathbf{a}) = \sum_{\text{all states}} \mathrm{p}(i)\ \log_2 [\mathrm{p}(i)]^{-1}$$

Following the law of exponents for logarithms, exponent –1 is eliminated by writing the equation as:

$$H(\mathbf{a}) = -\sum_{\text{all states}} \mathrm{p}(i) \log_2 \mathrm{p}(i)$$

Bit
Hartley
Decit
Nat

This is Shannon's formula for entropy (eq. 6.1). When the base for the logarithms is 2, the model is that of binary questions and the unit of entropy is the *bit* (contraction of *binary digit*) or *hartley* (Pinty & Gaultier, 1971). The model may be reformulated using questions with 10 answers, in which case the base of the logarithms is 10 and the unit is the *decit*. For natural logarithms, the unit is the *nat* (Pielou, 1975). These units are dimensionless, as are angles for example (Chapter 3).

Communication

Equation 6.1 may be applied to *human communications*, to calculate the information content of strings of symbols. For example, in a system of numbers with base n, there are n^N possible numbers containing N digits (in a base-10 system, there are $10^2 = 100$ numbers containing 2 digits, i.e. the numbers 00 to 99). It follows that the information content of a number with N digits is:

$$H = \log_2 n^N = N \log_2 n$$

The information per symbol (digit) is thus:

$$H/N = \log_2 n \qquad \textbf{(6.2)}$$

In the case of a binary (base 2) number, the information per symbol is $\log_2 2 = 1$ bit; for a decimal (base 10) number, it is $\log_2 10 = 3.32$ bits. A decimal digit contains 3.32 bits of information so that, consequently, a *binary* representation requires on average 3.32 times more digits than a *decimal* representation of the same number.

Alphabet

For an alphabet possessing 27 symbols (26 letters and the blank space), the information per symbol is $\log_2 27 = 4.76$ bits, assuming that all symbols have the same frequency. In languages such as English and French, each letter has a frequency of its own, so that the information per symbol is less than 4.76 bits. The information per letter is 4.03 bits in English and 3.95 bits in French. Hence, the translation from French to English should entail shorter text, which is generally the case.

English
French

Each language is characterized by a number of properties, such as the frequencies of letters, groups of letters, etc. These statistical properties, together with a defined syntax, determine a particular structure. For a given alphabet, the specific constraints

Table 6.3 Redundancy in the French language. Number of lexical elements with 4 to 6 letters (from Bourbeau *et al*., 1984).

Number of letters	Possible number of lexical elements	Actual number of lexical elements in French
4	$26^4 \approx$ 457 000	3 558
5	$26^5 \approx$ 12 000 000	11 351
6	$26^6 \approx$ 300 000 000	24 800

of a language limit the number of messages that can actually be formulated. Thus, the number of lexical elements with 4, 5 or 6 letters is much smaller than the theoretical possible number (Table 6.3). This difference arises from the fact that every language contains a certain amount of information that is inherently embodied in its structure, which is termed *redundancy*. Without redundancy, it would be impossible to detect errors slipping into communications, since any possible group of symbols would have meaning.

In a language with n different symbols, each having a characteristic frequency ($N_1, N_2 \ldots N_n$), the total number of possible messages (P) made up of N symbols is equal to the number of *combinations*:

$$P = N! / (N_1!\, N_2! \ldots N_n!)$$

The information content of a message with N symbols is:

$$H = \log_2 P = \log_2 [N! / (N_1!\, N_2! \ldots N_n!)]$$

Hence, the information per symbol is:

$$H/N = 1/N \log_2 [N! / (N_1!\, N_2! \ldots N_n!)] \qquad \textbf{(6.3)}$$

which is the formula of Brillouin (1956). It will be used later (Section 6.5) to calculate the species diversity of a sample, considered to be representing a "message".

6.2 Two-way contingency tables

In order to compare two qualitative descriptors, the objects are allocated to the cells of a table with two criteria, i.e. the rows and columns. Each cell of the *two-way contingency table* (e.g. Tables 6.1 and 6.4) contains the number of observations

Table 6.4 Contingency table giving the observed (from Table 6.1) and expected (in parentheses) frequencies in each cell; $n = 120$. The observed frequencies that exceed the corresponding expected frequencies are in boldface. Wilks' chi-square statistic: $X^2_W = 150.7$ ($\nu = 9, p < 0.001$).

	b_1	b_2	b_3	b_4
	30	30	30	30
$a_1 = 60$	**30** (15)	10 (15)	15 (15)	5 (15)
$a_2 = 30$	0 (7.5)	**20** (7.5)	0 (7.5)	**10** (7.5)
$a_3 = 15$	0 (3.75)	0 (3.75)	0 (3.75)	**15** (3.75)
$a_4 = 15$	0 (3.75)	0 (3.75)	**15** (3.75)	0 (3.75)

described by that pair of states of the qualitative descriptors. Numbers in the cells of a contingency table are absolute frequencies, i.e. *not* relative frequencies. The number of cells in the table is equal to the product of the number of states in the two descriptors. The first question relative to a contingency table concerns the relationship between the two descriptors: given the bivariate distribution of observations in the table, are the two descriptors related to each other, or not? This question is answered by calculating the expected frequency E for each cell of the table, according to a null hypothesis H_0, and performing a chi-square (X^2) test of the null hypothesis.

Null hypothesis

The simplest null hypothesis is the independence of the two descriptors. E_{ij} is the number of observations that is expected in each cell (i, j) under H_0. Under this null hypothesis, E_{ij} is computed as the product of the marginal totals (i.e. the product of the sum of row i with the sum of column j), divided by n which is the total number of observations in the table:

Expected frequency

$$E_{ij} = [(\text{row sum})_i \times (\text{column sum})_j] / n \qquad \textbf{(6.4)}$$

This equation generates expected frequencies whose relative distribution across the states of descriptor **a**, *within* each state of descriptor **b**, is the same as the distribution of all observed data across the states of **a**, and conversely (Table 6.4). The null hypothesis is tested using a X^2-statistic that compares the observed (O_{ij}) to the expected frequencies (E_{ij}).

In basic statistics textbooks, the significance of relationships in two-way contingency tables is often tested using the *Pearson chi-square statistic* (Pearson, 1900):

Pearson chi-square

$$X_P^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} \qquad \textbf{(6.5)}$$

where $(O - E)$ measures the contingency of each cell. Instead of X_P^2, one can compute Wilks' likelihood ratio (1935), also known as the *G* or *2I-statistic* (Sokal & Rohlf, 1995) or G^2 (Bishop *et al*., 1975; Dixon, 1981):

Wilks chi-square

$$X_W^2 = 2 \sum_{\text{all cells}} O \log_e\left(\frac{O}{E}\right) \qquad \textbf{(6.6)}$$

where $\log_e$ is the natural logarithm. For null frequencies, $\lim_{O \to 0} [O \log_e (O/E)] = 0$.

Degrees of freedom

For a contingency table with *r* rows and *c* columns, the number of degrees of freedom used to determine the probability (p-value) of the data under H_0 is:

$$\nu = (r-1)(c-1) \qquad \textbf{(6.7)}$$

When the p-value is smaller than or equal to a predetermined significance level, e.g. $\alpha = 0.05$, the null hypothesis (H_0) of independence of the descriptors is rejected.

When the number of observations (*n*) is large (i.e. larger than ten times the number of cells, *rc*, in the table), the asymptotic distributions of X_P^2 and X_W^2 are both χ^2. In other words, the two statistics are equivalent when H_0 is true. There is however a problem when the number of observations is small, i.e. less than five times the number of cells. Small numbers of observations often lead to several null observed values (O_{ij}) in the contingency table, with correspondingly very low expected frequencies (E_{ij}). According to Cochran (1954) and Siegel (1956), when there is at least *one* value of E_{ij} smaller than 1, or when 20% or more of the expected values E_{ij} are smaller than 5, some states (rows or columns) must be grouped to increase the expected frequencies, provided that there is a logical basis to do so. It now appears that only the first part of this empirical rule should be kept. Indeed Fienberg (1980, p.172) cites results of simulations indicating that, for $\alpha = 0.05$, the computed statistic is distributed like χ^2 if H_0 is true, as long as all E_{ij} values are larger than 1.

Concerning the choice of X_P^2 or X_W^2, there is no difference when the number of observations *n* is large (see the previous paragraph). When *n* is small, Larntz (1978) is of the opinion that X_P^2 is better than X_W^2. Sokal & Rohlf (1995) still recommend using X_W^2 but suggest to correct it as proposed by Williams (1976a) to obtain a better approximation of χ^2. This correction consists in dividing X_W^2 by a correction factor q_{min}. The correction factor, which is based on ν (eq. 6.7), is computed as:

Williams' correction

$$q_{min} = 1 + [(r^2-1)(c^2-1)/6\nu n] \qquad \textbf{(6.8)}$$

When n is large relative to the number of cells in the contingency table, it is not necessary to apply a correction to X^2_W since $q_{min} \approx 1$ in that case. William's correction is especially interesting when one must use X^2_W, as in the study of multiway contingency tables; the general formula for q_{min} is given in Subsection 6.3. Several computer programs allow users to compute both X^2_P and X^2_W.

Another correction, available in some computer programs, consists in adding a small value (e.g. 0.5) to *each* observed value O_{ij} in the contingency table when some of the O_{ij}'s are small. As indicated by Dixon (1981) and Sokal & Rohlf (1995), the effect of this correction is to lower the X^2-statistic, which makes the test more conservative. H_0 may then be rejected in a proportion of cases smaller than α when the null hypothesis is true.

Another measure of interest to ecologists, which is related to the Wilks statistic (see below), refers to the concept of entropy (or information) discussed above. In the numerical example with four rows and columns (Tables 6.1 and 6.4), if the correspondence between the states of descriptors **a** and **b** was perfect (i.e. descriptors completely dependent of each other), the contingency table would only have four non-zero cells — one in each row and each column. These non-zero cells could be anywhere in the table, not necessarily on the diagonal, because the states of the two descriptors are not ordered. It would then be possible, using **a**, to perfectly predict the distribution of observations among the states of **b**, and vice versa. In other words, given one state of the first descriptor, one would immediately know the state of the other descriptor. Thus, there would be no uncertainty (or entropy) concerning the distribution of the objects on **b** after observing **a**, hence the entropy remaining in **b** after observing **a** would be null, i.e. $H(\mathbf{b}|\mathbf{a}) = 0$. On the contrary, if the descriptors were completely independent of each other, the distribution of observations in each row of descriptor **a** would be in the same proportions as their overall distribution in **b** (found at top of Tables 6.1 and 6.4); the same would be true for the columns. $H(\mathbf{b}|\mathbf{a}) = H(\mathbf{b})$ would indicate that all the entropy contained in the distribution of **b** remains after observing **a**.

The two conditional entropies $H(\mathbf{a}|\mathbf{b})$ and $H(\mathbf{b}|\mathbf{a})$, as well as the entropy shared by the two descriptors, can be computed using the total information in the contingency table, $H(\mathbf{a},\mathbf{b})$, and the information of each descriptor, $H(\mathbf{a})$ and $H(\mathbf{b})$, already computed in Section 6.1. $H(\mathbf{a},\mathbf{b})$ is computed on all observed frequencies in the contingency table using Shannon's formula (eq. 6.1):

$$H(\mathbf{a},\mathbf{b}) = -\sum_{\text{states of } \mathbf{a}} \sum_{\text{states of } \mathbf{b}} p(i,j) \log p(i,j) \qquad \textbf{(6.9)}$$

where $p(i,j)$ is the observed frequency in each cell (i,j) of the contingency table, divided by the total number of observations n. For the example (Tables 6.1 or 6.4):

$$H(\mathbf{a},\mathbf{b}) = -\{1/4 \log_2 (1/4) + 1/6 \log_2 (1/6) + 3\,[1/8 \log_2 (1/8)] + 2\,[1/12 \log_2 (1/12)] + 1/24 \log_2 (1/24)\} = 2.84$$

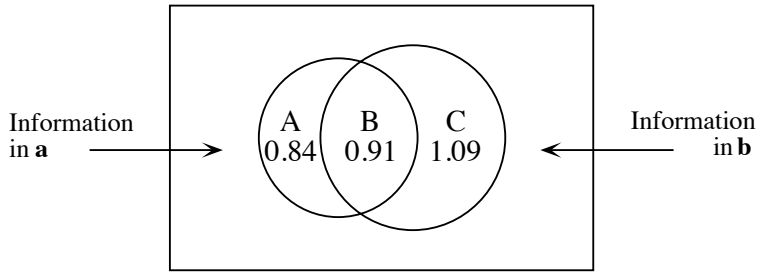


Figure 6.1 Venn diagram partitioning the information of two qualitative descriptors, denoted **a** and **b**. B is the information the two descriptors have in common.

The values of $H(\mathbf{a}) = A + B = 1.75$ and $H(\mathbf{b}) = B + C = 2.00$, represented by circles in the Venn diagram of Fig. 6.1, have been computed in Section 6.1. $H(\mathbf{a},\mathbf{b}) = 2.84$ is the total information in the union of the two descriptors, represented by A + B + C. The information (B) shared by the two descriptors is computed as follows:

$$B = (A + B) + (B + C) - (A + B + C)$$

$$B = H(\mathbf{a}) + H(\mathbf{b}) - H(\mathbf{a},\mathbf{b}) \qquad \textbf{(6.10)}$$

$$B = 1.75 + 2.00 - 2.84 = 0.91$$

With more decimals, B = 0.90564; this value is used in the example that follows eq. 6.14. The information exclusive to each descriptor, A and C, is computed by subtraction as follows:

$$A = (A + B + C) - (B + C)$$

$$A = H(\mathbf{a}|\mathbf{b}) = H(\mathbf{a},\mathbf{b}) - H(\mathbf{b}) \qquad \textbf{(6.11)}$$

$$A = 2.84 - 2.00 = 0.84$$

and

$$C = (A + B + C) - (A + B)$$

$$C = H(\mathbf{b}|\mathbf{a}) = H(\mathbf{a},\mathbf{b}) - H(\mathbf{a}) \qquad \textbf{(6.12)}$$

$$C = 2.84 - 1.75 = 1.09$$

There is a relationship between the reciprocal information B and Wilks X^2_W statistic. It can be shown that $B = (1/n) \sum O \log_e(O/E)$ when B is computed with natural logarithms ($\log_e$), or else $B \log_e 2 = (1/n) \sum O \log_e(O/E)$ when B is in bits. Using these relationships, it is possible to calculate the probability associated with B after transforming B into a Wilks X^2_W-statistic (eq. 6.6):

$$X^2_W = 2nB \quad \text{when B is in nats} \qquad \textbf{(6.13)}$$

or $X^2_W = 2n\mathrm{B}\log_e 2 = n\mathrm{B}\log_e 4 = 1.38629\, n\mathrm{B}$ when B is in bits. **(6.14)**

For the numerical example, $X^2_W = 2n\mathrm{B}\log_e 2 = 2 \times 120 \times 0.90564 \times 0.69315 = 150.66$ before Williams' correction.

Similarity

Using the measures of information A, B and C, various reciprocal information coefficients can be computed. The *similarity* of descriptors **a** and **b** can be calculated as the amount of information that the two descriptors have in common, divided by the total information of the system:

$$S(\mathbf{a},\mathbf{b}) = \mathrm{B} / (\mathrm{A} + \mathrm{B} + \mathrm{C}) \quad \textbf{(6.15)}$$

$S(\mathbf{a},\mathbf{b}) = 0.91 / 2.84 = 0.32$, for the numerical example.

If the following steps of the analysis (clustering and ordination, Chapters 8 and 9) require that the measure of association between **a** and **b** be a metric, one may use the corresponding distance, defined as the sum of the information that the two descriptors possess independently, divided by the total information:

Rajski's metric

$$D(\mathbf{a},\mathbf{b}) = (\mathrm{A} + \mathrm{C}) / (\mathrm{A} + \mathrm{B} + \mathrm{C}) \quad \textbf{(6.16)}$$

For the numerical example, $D(\mathbf{a},\mathbf{b}) = (0.84 + 1.09) / 2.84 = 0.68$. As indicated by the structure of the formulas, $S(\mathbf{a},\mathbf{b}) + D(\mathbf{a},\mathbf{b}) = 1$.

Coherence coefficient

The distance measure in eq. 6.16 is Rajski's metric (1961). This author also proposed another measure of similarity among descriptors, the *coherence coefficient*, which is used to assess the stochastic independence of two random variables:

$$S' = \sqrt{1 - D^2} \quad \textbf{(6.17)}$$

Another version of this coefficient,

$$\mathrm{S}'' = \mathrm{B} / (\mathrm{A} + 2\mathrm{B} + \mathrm{C}) \quad \textbf{(6.18)}$$

Symmetric, asymmetric uncertainty

is available in some computer programs under the name *symmetric uncertainty coefficient*. Two *asymmetric uncertainty coefficients* have also been proposed. They are used, for example, to compare the explanatory power of a given descriptor with respect to several other descriptors: B / (A + B) controls for the total amount of information in **a**, whereas B / (B + C) controls for the total information in **b**.

The construction of an association matrix, containing any of the symmetric coefficients described above, requires calculating $p(p-1)/2$ contingency tables; this matrix is symmetric and its diagonal is $S = 1$ or $D = 0$. *Qualitative (nonordered) descriptors* can thus be used to compute *quantitative association coefficients*, which makes possible the numerical analysis of multivariate qualitative data sets. Furthermore, since quantitative or semiquantitative descriptors can be recoded into

discrete states, it is possible, using uncertainty coefficients, to compute association matrices among descriptors of mixed types.

It is only through B, which can be transformed into a X^2_W-statistic, that a probability can be associated to the uncertainty coefficients. For coefficient S above (eq. 6.15), short of computing a p-value, one can state in general terms that two descriptors are very closely related when $S(\mathbf{a},\mathbf{b}) > 0.5$; they are well associated when $0.5 > S > 0.3$; and some association exists when $S < 0.3$ without coming too close to 0 (Hawksworth *et al*., 1968).

The probability associated with a X^2-statistic, calculated on a contingency table, assesses the hypothesis that the relationship between the two descriptors is *random*. Biological associations, for example, could be defined on the basis of relationships found to be non-random between pairs of species — the relationship being defined by reference to a pre-selected probability level (e.g. $\alpha = 0.05$ or 0.01) associated with the X^2 measuring the contingency between two species (Subsection 7.5.2). The value of X^2 may itself be used as a measure of the *strength* of the relationship between species. This is also the case for the reciprocal information measures defined above. With the same purpose in mind, it is possible to use one of the following *contingency coefficients*, which are merely transformations of a X^2-statistic on a scale from 0 to 1 (Kendall & Buckland, 1960; Morice, 1968):

Contingency coefficient

$$\text{Pearson contingency coefficient}, C = \sqrt{X^2 / (n + X^2)} \qquad \textbf{(6.19)}$$

$$\text{Tschuproff contingency coefficient}, T = \sqrt{X^2 / (n\sqrt{\text{degrees of freedom}})} \qquad \textbf{(6.20)}$$

where n is the number of observations. These contingency coefficients are not frequently used in ecology, however. They can only be used for comparing contingency tables of the same sizes.

Contingency tables are the main approach available to ecologists for the numerical analysis of relationships among qualitative descriptors, or else between qualitative descriptors and ordered variables divided into classes. Contingency tables are also convenient for analysing *nonmonotonic relationships* among ordered descriptors (a relationship is monotonic when there is a constant evolution of a descriptor with respect to the other; see Fig. 5.1). Reciprocal information and X^2 coefficients are sensitive enough that they could be used even with ordered variables, when relationships among a large number of descriptors are analysed by computer. One must simply make sure that the ordered data are divided into a sufficiently large number of classes to avoid clumping together observations that one would want to keep distinct in the results. If a first analysis indicates that redefining the boundaries of the classes could improve the interpretation of the phenomenon under study (the classes used to recode quantitative variables do not need to have the same width), ecologists should not hesitate to repeat the analysis using the recoded data. This procedure is not circular; it corresponds to a progressive discovery of the structure of the information.

It is also possible to use the association coefficients described above to interpret the classifications resulting from a first analysis of the data (Chapter 8). A classification may be compared to the descriptors from which it originates, in order to determine which descriptors are mostly responsible for it; or else, it may be compared to a new series of descriptors that could potentially explain it. One can also use contingency tables to compare several classifications of the same objects, obtained through different methods. Subsection 10.2.1 describes these higher-level analyses.

Ecological application 6.2

Legendre *et al*. (1978) analysed data from a winter aerial survey of land fauna, using contingency tables. They compared the presence or absence of tracks of different bird and mammal species to a series of 11 environmental descriptors. Five of these descriptors were qualitative, i.e. bioclimatic region, plant association, nature of the dominant and sub-dominant surface materials, and category of aquatic ecosystem. The others were semiquantitative, i.e. height of the trees, drainage, topography, thickness of the surface materials, abundance of streams and wetlands. The analysis identified the descriptors that determined or limited the presence of the 10 species that had been observed with sufficient frequency to permit their analysis. This allowed the authors to describe the niches of these species.

6.3 Multiway contingency tables

When there are more than two descriptors, one might consider the possibility of analysing the data set using a series of two-way contingency tables, in which each pair of descriptors would be treated separately. Such an approach, however, would not take into account possible interactions among several descriptors and might thus miss part of the potential offered by the multidimensional structure of the data. This could lead to incorrect, or at least incomplete conclusions. Information on the analysis of multiway contingency tables can be found in Kullback (1959), Plackett (1974), Bishop *et al*. (1975), Upton (1978), Gokhale & Kullback (1978), Fienberg (1980), Sokal & Rohlf (1995), Agresti (2002), and Kroonenberg (2008).

Log-linear model

The most usual approach for analysing multiway contingency tables is to adjust to the data a *log-linear model*, where the natural logarithm ($\log_e$) of the expected frequency E for each cell of the table is estimated as a sum of main effects and interactions. For example, in the case of two-way contingency tables (Section 6.2), the expected frequencies could have been computed using the following equation:

$$\log_e E = [\theta] + [A] + [B] + [AB] \qquad \textbf{(6.21)}$$

Symbols in brackets are the *effects*. [A] and [B] are the main effects of descriptors **a** and **b**, respectively, and [AB] is the effect resulting from the interaction between **a** and **b**. [θ] is the mean of the logarithms of the expected frequencies. In a two-way table,

the hypothesis tested is that of independence between the two descriptors, i.e. H_0: [AB] = 0. The log-linear model corresponding to this hypothesis is thus:

$$\log_e E = [\theta] + [A] + [B] \quad \textbf{(6.22)}$$

since [AB] = 0. The expected frequencies E computed using eq. 6.22 are exactly the same as those computed in Section 6.2 (eq. 6.4). Hence for two-way tables, one usually computes the expected frequencies with eq. 6.4. For multiway tables, the expected frequencies are generated with an iterative proportional fitting algorithm. The advantage of log-linear models is obvious when analysing contingency tables with more than two dimensions (or *criteria*).

For a contingency table with three descriptors (**a**, **b**, and **c**), the log-linear model containing all possible effects is:

$$\log_e E = [\theta] + [A] + [B] + [C] + [AB] + [AC] + [BC] + [ABC]$$

Saturated model

Such a model is referred to as the *saturated model*. In practice, the effect resulting from the interaction among all descriptors is never included in any log-linear model, i.e. here [ABC]. This is because the expected frequencies for the saturated model are always equal to the observed frequencies ($E = O$), so that this model is useless. The general log-linear model for a three-way table is thus:

$$\log_e E = [\theta] + [A] + [B] + [C] + [AB] + [AC] + [BC] \quad \textbf{(6.23)}$$

where H_0: [ABC] = 0. In other words, the logarithm of the expected frequency for each cell of the contingency table is computed here by adding, to the mean of the logarithms of the expected frequencies, one effect due to each of the three descriptors and one effect resulting from each of their two-way interactions.

Hierarchical model

Different log-linear models may be formulated by setting some of the effects equal to zero. Normally, one only considers *hierarchical models*, i.e. models in which the presence of a higher-order effect implies that all the corresponding lower effects are also included; the order of an effect is the number of symbols in the bracket. For example, in a hierarchical model, to include [BC] implies that both [B] and [C] are also included. For a three-way contingency table, there are eight possible hierarchical models, corresponding to as many different hypotheses (Table 6.5). Models in the table all include the three main effects. Each hypothesis corresponds to different types of interaction among the three variables. In practice, one uses a program available in a computer package (for R functions, see Section 6.6), with which it is easy to estimate the expected frequencies for any hierarchical model of interest to the user.

The number of degrees of freedom (ν) depends on the interactions that are included in the model. For the general hierarchical model of eq. 6.23,

$$\nu = rst - [1 + (r-1) + (s-1) + (t-1) + (r-1)(s-1) + (r-1)(t-1) + (s-1)(t-1)] \quad \textbf{(6.24)}$$

Table 6.5 Possible log-linear models for a three-way contingency table. Hypotheses and corresponding models. All models include the three main effects [A], [B] and [C].

Hypotheses (H_0)	Log-linear models
1.[ABC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]+[AC]+[BC]
2.[ABC] = 0, [AB] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AC]+[BC]
3.[ABC] = 0, [AC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]+[BC]
4.[ABC] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]+[AC]
5.[ABC] = 0, [AB] = 0, [AC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[BC]
6.[ABC] = 0, [AB] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AC]
7.[ABC] = 0, [AC] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]+[AB]
8.[ABC] = 0, [AB] = 0, [AC] = 0, [BC] = 0	$\log_e E$ = [θ]+[A]+[B]+[C]

where r, s and t are the numbers of states of descriptors **a**, **b** and **c**, respectively. If there were only two descriptors, **a** and **b**, the log-linear model would not include the interaction [AB], so that eq. 6.24 would become:

$$\nu = rs - [1 + (r - 1) + (s - 1)] = (r - 1)(s - 1)$$

which is identical to eq. 6.7. In Table 6.5, model 4, for example, does not include the interaction [BC], so that:

$$\nu = rst - [1 + (r - 1) + (s - 1) + (t - 1) + (r - 1)(s - 1) + (r - 1)(t - 1)]$$

Programs in computer packages calculate the number of degrees of freedom corresponding to each model.

It is possible to test the goodness of fit of a given model to the observed data by using one of the X^2 statistics already described for two-way tables, X^2_P or X^2_W (eqs. 6.5 and 6.6). The null hypothesis (H_0) tested is that the effects excluded from the model are null. Rejecting H_0, however, does not allow one to accept the alternative hypothesis that *all* the effects included in the model are not null. The only conclusion to be drawn from rejecting H_0 is that at least some of the effects in the model are not null. When the p-value associated with a model is larger than the significance level α, the conclusion is that the model fits the data well.

Williams' correction

As in the case of two-way contingency tables (eq. 6.8), it is recommended to divide X^2_W by a correction factor, q_{min} (Williams, 1976a), when the number of observations n is small, i.e. less than 4 or 5 times the number of cells in the table. For the general hierarchical model (eqs. 6.23 and 6.24):

$$q_{min} = 1 + (1/6\nu n)\,[r^2s^2t^2 - 1 - (r^2 - 1) - (s^2 - 1) - (t^2 - 1) - (r^2 - 1)(s^2 - 1) - (r^2 - 1)(t^2 - 1) - (s^2 - 1)(t^2 - 1)] \qquad \textbf{(6.25)}$$

In the case of two descriptors, eq. 6.25 becomes:

$$q_{min} = 1 + (1/6\nu n)\,[r^2s^2 - 1 - (r^2 - 1) - (s^2 - 1)]$$

$$q_{min} = 1 + (1/6\nu n)\,[(r^2 - 1)(s^2 - 1)$$

which is identical to eq. 6.8. For model 4 in Table 6.5, used above as example:

$$q_{min} = 1 + (1/6\nu n)\,[r^2s^2t^2 - 1 - (r^2 - 1) - (s^2 - 1) - (t^2 - 1) - (r^2 - 1)(s^2 - 1) - (r^2 - 1)(t^2 - 1)]$$

This correction cannot be applied, as such, to contingency tables containing null expected frequencies (see below). The other possible correction, which consists in adding to each cell of the table a small value, e.g. 0.5, has the same effect here as in two-way contingency tables (see Section 6.2).

Ecological application 6.3a

Legendre (1987a) analysed biological oceanographic data obtained at 157 sites in Baie des Chaleurs (Gulf of St. Lawrence, eastern Canada). The data set (observations made at 5-m depth) included measurements of temperature, salinity, nutrients (phosphate and nitrate), and chlorophyll *a* (estimated from the *in vivo* fluorescence of water pumped on board the ship). As it often happens in ecology, the numerical analysis was hampered by three practical problems. (1) The measured concentrations of nutrients were often near or below the detection limit, with the result that many of them exhibited large experimental errors (since the 1980s, the detection limits of some nutrients have been lowered by a factor 100 or 1000). (2) Relationships between variables were often nonmonotonic, i.e. they did not continuously increase or decrease but reached a maximum (or a minimum) after which they decreased (or increased). (3) Most of the variables were intercorrelated, so that no straightforward interpretation of phytoplankton (i.e. chlorophyll *a*) concentrations was possible in terms of environmental variables. Since multiway contingency table analysis can handle these three types of problems, it was decided to partition the (ordered) variables into discrete classes and analyse the transformed data using hierarchical log-linear models.

The initial model in Table 6.6 (line 1) only included the interaction among the three environmental variables, with no effect of these on chl *a*. This model did not fit the data well. Adding the interaction between chlorophyll *a* (chl *a*) and the temperature-salinity (TS) characteristics significantly improved the fit (i.e. there was a significant difference between models; line 2). The resulting model could be accepted (line 3), but adding the interaction between chl *a* and phosphate further improved the fit (significant difference, line 4) and the

Table 6.6 Multiway contingency table analysis of oceanographic data recoded into discrete classes (Legendre, 1987a). Using a hierarchy of log-linear models, the concentrations of chlorophyll *a* (symbol in this table: C, 4 classes) are analysed as a function of the temperature-salinity (TS) characteristics of the water masses (symbol in this table: T, 3 classes) and the concentrations of phosphate (P; 2 classes) and nitrate (N; 2 classes). When a higher-order effect is present, all the corresponding lower-order effects are included in the model.

Effects in the model	Interpretation	ν	X^2_W
[NTP], [C]	Chl *a* is independent of the environmental variables	30	121 *
Difference	Adding [CT] to the model significantly improves the fit	9	89 *
[NTP], [CT]	Chl *a* depends on the TS characteristics	21	**32**
Difference	Adding [CP] to the model significantly improves the fit	3	13 *
[NTP], [CT], [CP]	Chl *a* depends on the TS characteristics and on phosphate	18	**19**
Difference	Adding [CN] does not significantly improve the fit	7	5
[NTP], [CT], [CP], [CN]	The most parsimonious model does not include a dependence of chl *a* on nitrate	11	**14**

* $p \leq 0.05$; bold X^2_W values correspond to models with $p > 0.05$ of fitting the data

resulting model fitted the data well (line 5). Final addition of the interaction between chl *a* and nitrate did not improve the fit (difference not significant, line 6). The most parsimonious model (line 5) thus showed a dependence of chl *a* concentrations on the TS characteristics and phosphate. The choice of the initial model in Table 6.6 is explained in Ecological application 6.3b.

There are 8 hierarchical models associated with a three-way contingency table, 113 with a four-way table, and so forth, so that the choice of a single model, among all those possible, rapidly becomes a major problem. In fact, it often happens that several models fit the data well. Also, in many instances, the fit to the data could be improved by adding supplementary terms (i.e. effects) to the model. However, this improved fit would result in a more complex ecological interpretation because of the added interaction(s) among descriptors. It follows that the choice of a model generally involves a compromise between goodness of fit and simplicity of interpretation, as suggested by the principle of parsimony (Subsection 10.3.3). Finally, even when it is possible to test the fit of all possible models to the data, this way of proceeding involves multiple testing and the p-values require correction (Box 1.3).

To select a model, there are several methods that are both statistically acceptable and ecologically parsimonious. In practice, because no method is totally satisfactory, one could simply use, with care, those included in the available computer package.

Partitioning the X^2_W

1. A first method consists in *partitioning* the X^2_W statistics associated with a hierarchy of log-linear models. The hierarchy contains a series of models, which are made progressively simpler (or more complex) by removing (or adding) one effect at a time. It can be shown that the difference between the X^2_W statistics of two successive models in the hierarchy is itself a X^2_W-statistic, which can therefore be tested. The corresponding number of degrees of freedom is the difference between those of the two models. This is the approach used in Ecological application 6.3a (see Table 6.6). The main problem with this method is that one may find different "most parsimonious" models depending on the hierarchy chosen *a priori*. Partitioning X^2 statistics is possible only with X^2_W, not X^2_P.

Stepwise selection

2. A second family of approaches lies in the *stepwise forward selection* or *backward elimination* of terms in the model. As always with stepwise methods (see Subsection 10.3.3), (a) it may happen that forward selection lead to models quite different from those resulting from backward elimination, and (b) the tests of significance must be interpreted with caution because the computed statistics are not independent. Stepwise methods thus only provide guidance, which may be used for limiting the number of models to be considered. It often happens that models other than those identified by the stepwise approach are found to be more parsimonious and interesting, and to fit the data just as well (Fienberg, 1980: 80).

Effect screening

3. Other methods simultaneously consider all possible effects. An example of *effect screening* (Brown, 1976) is given in Dixon (1981). The approach is useful for reducing the number of models to be subsequently treated, for example, by the method of hierarchical partitioning of X^2_W statistics (see method 1 above).

When analysing multiway contingency tables, ecologists must be aware of a number of possible practical problems, which may sometimes have significant impact on the results. These potential problems concern the cells with zero expected frequencies, the limits imposed by the sampling design, the simultaneous analysis of descriptors with mixed levels of precision (i.e. qualitative, semiquantitative, and quantitative), and the use of contingency tables for the purpose of explanation or forecasting.

Cells with $E = 0$

1. Multiway contingency tables, in ecology, often include cells with expected frequencies $E = 0$. There are two types of zero expected frequencies, i.e. those resulting from sampling and those that are of structural nature.

Sampling zeros are caused by random variation, combined with small sample size relative to the number of cells in the multiway contingency table. Such zeros would normally disappear if the size of the sample was increased. The presence of cells with null observations ($O = 0$) may result, when calculating specific models, in some

expected frequencies $E = 0$. This is accompanied by a reduction in the number of degrees of freedom. For example, according to eq. 6.24, the number of degrees of freedom for the initial model in Table 6.6 (line 1) should be $\nu = 33$, since this model includes four main effects [C], [N], [P], and [T] and interactions [NP], [NT], [PT], and [NPT]; however, the presence of cells with null observations ($O = 0$) led to cells with $E = 0$, which reduced the number of degrees of freedom to $\nu = 30$. Rules to calculate the reduction in the number of degrees of freedom are given in Bishop *et al*. (1975: 116 *et seq*.) and Dixon (1981: 666). In practice, computer programs generally take into account the presence of zero expected frequencies when computing the number of degrees of freedom for multiway tables. The problem does not occur with two-way contingency tables because cells with $E = 0$ are only possible, in the two-way configuration, if all the observations in the corresponding row or column are null, in which case the corresponding state is automatically removed from the table.

Structural zeros correspond to combinations of states that cannot occur *a priori* or by design. For example, in a study where two of the descriptors are sex (female, male) and sexual maturity (immature, mature, gravid), the expected frequency of the cell "gravid male" would *a priori* be $E = 0$. Another example would be combinations of states that have not been sampled, either by design or involuntarily (e.g. lack of time, or inadequate planning). Several computer programs allow users to specify the cells that contain structural zeros, before computing the expected frequencies.

2. In principle, the methods described here for multiway contingency tables can only be applied to data resulting from *simple random sampling* or *stratified sampling* designs. Fienberg (1980: 32) gives some references in which methods are described for analysing qualitative descriptors within the context of *nested sampling* or a *combination of stratified and nested sampling* designs. Sampling designs are described in Cochran (1977), Green (1979), and Thompson (1992), for example.

Mixed precision

3. Analysing together *descriptors with mixed levels of precision* (e.g. a mixture of qualitative, semiquantitative, and quantitative descriptors) may be done using multiway contingency tables. In order to do so, continuous descriptors must first be partitioned into a small number of classes. Unfortunately, there exists no general approach to do so. When there is no specific reason for setting the class limits, it has been suggested, for example, to partition continuous descriptors into classes of equal width, or containing an equal number of observations. Alternatively, Cox (1957) describes a method that may be used for partitioning a normally distributed descriptor into a predetermined number of classes (2 to 6). For the specific case discussed in the next paragraph, where there is one response variable and several explanatory variables, Legendre & Legendre (1983b) describe a method for partitioning the ordered explanatory variables into classes in such a way as to maximize the relationships to the response variable. It is important to be aware that, when analysing the contingency table, different ways of partitioning continuous descriptors may sometimes lead to different conclusions. In practice, the number of classes of each descriptor should be as small as possible, in order to minimize the problems discussed above concerning the calculation of X^2_W (see eqs. 6.8 ad 6.25 for correction factor q_{min}) and the presence of

sampling zeros. Another point is that contingency table analysis considers the different states of any descriptor to be nonordered. When some of the descriptors are in fact ordered (i.e. originally semiquantitative or quantitative), the information pertaining to the ordering of states may be used when adjusting log-linear models (see for example Fienberg, 1980: 61 *et seq*.).

4. There is an analogy between *log-linear models* and *analysis of variance* since the two approaches use the concepts of effects and interactions. This analogy is superficial, however, since analysis of variance aims at assessing the effects of explanatory factors on a single response variable, whereas log-linear models have been developed to describe structural relationships among several descriptors corresponding to the dimensions of the table.

5. It is possible to use contingency table analysis for interpreting a *response variable* in terms of several interacting *explanatory variables*. In such a case, the following basic rules must be followed. (1) Any log-linear model fitted to the data must include by design the term for the highest-order interaction among all *explanatory variables*. In this way, all possible interactions among the explanatory variables are included in the model, because of its hierarchical nature. (2) When interpreting the model, one should not discuss the interactions among the explanatory variables. They are incorporated in the model for the reason given above, but no test of significance is performed on them. In any case, one is only interested in the interactions between the explanatory and response variables. An example follows.

Ecological application 6.3b

The example already discussed in application 6.3a (Legendre, 1987a) aimed at interpreting the horizontal distribution of phytoplankton in Baie des Chaleurs (Gulf of St. Lawrence, eastern Canada) in terms of selected environmental variables. In such a case, where a single response variable is interpreted as a function of several potentially explanatory variables, all models considered must include by design the highest-order interaction among the explanatory variables. Thus, all models in Table 6.6 included the interaction [NPT]. The simplest model in the hierarchy (line 1 in Table 6.6) only contained [NPT] and [C] as effects. In this simplest model, there was no interaction between chlorophyll and any of the three environmental variables, i.e. the model did not include [CN], [CP] or [CT]. When interpreting the model selected as best fitting the data, the author did not discuss the interaction among the explanatory variables because the presence of [NPT] prevented a proper analysis of this interaction. Table 6.6 then led to the interpretation that the horizontal distribution of phytoplankton depended on the TS characteristics of water masses and phosphate concentration.

Logistic regression

When the *qualitative response variable* is *binary*, one may use the *logistic linear* (or *logit*) *model* instead of the log-linear model. Fitting such a model to data is also called *logistic regression* (Subsection 10.3.7). In logistic regression, the explanatory descriptors do not have to be divided into classes; they may be discrete or continuous. This type of regression is available in various computer packages and in R (Section 10.7). Some programs allow the *response variable* to be *multi-state*. Efficient use of logistic regression requires that *all* the explanatory descriptors be potentially

related to the response variable. This method can replace discriminant analysis in cases discussed in Subsection 10.3.7 and Section 11.6.

Examples of successful use of multiway contingency tables in ecology include Fienberg (1970) and Schoener (1970) for the habitat of lizards, Jenkins (1975) for the selection of trees by beavers, Legendre & Legendre (1983b) for marine benthos, Fréchet (1990) for cod fishery, Schoener & Adler (1991) for spatial distributions of lizards and birds, Fedriani *et al*. (2001) for responses of coyote populations to anthropogenic food, Fingerut *et al*. (2003) for transmission of a marine parasite by swimming larvae, and Gorelick & Bertram (2010) for computation of diversity indices.

6.4 Contingency tables: correspondence

Once it has been established that two or more qualitative descriptors in a contingency table are not independent (Sections 6.2 and 6.3), it is often of interest to identify the cells of the table that account for the existing relationship between descriptors. These cells, which show how the descriptors are related, define the *correspondence* between the rows and columns of the contingency table. By comparison with parametric and nonparametric statistics (Chapters 4 and 5), the measures of contingency described in Sections 6.2 and 6.3 are, for qualitative descriptors, analogous to the *correlation* between ordered descriptors, whereas correspondence would be analogous to *regression* (Section 10.3) because it can be used to forecast the state of one descriptor using another descriptor. *Correspondence analysis* (Section 9.2) is another method that allows, among other objectives, the identification of the relationships between the rows and columns of a contingency table. This can be achieved directly through the approach described in the present section.

In a contingency table where the descriptors are not independent (i.e. the null hypothesis of independence has been rejected), the cells of interest to ecologists are those in which the observed frequencies (O_{ij}) are very different from the corresponding expected frequencies (E_{ij}). Each of these cells corresponds to a given state for each descriptor in the contingency table. The fact that $O_{ij} \neq E_{ij}$ is indicative of a stronger interaction, between the states in question, than expected under the null hypothesis which in invoked to compute E. For example, hypothesis H_0 in Table 6.4 is that of independence of descriptors **a** and **b**. This hypothesis having been rejected ($p < 0.001$), one may identify in the contingency table the observed frequencies O_{ij} that are much higher or lower than the corresponding expected frequencies E_{ij}. Values $O_{ij} > E_{ij}$ (in bold-face type in Table 6.4) give a preliminary indication of the associations between states of **a** and **b**. These values may be located anywhere in the table since contingency table analysis does not take into account the ordering of states.

When the test of the global X^2-statistic (eq. 6.5 or 6.6) supports the hypothesis of a significant relationship between the two descriptors, one can identify the cells that

Test of $O_{ij} = E_{ij}$

strongly contribute to the correspondence by testing the significance of the difference between O_{ij} and E_{ij} *in each cell* of the contingency table. Ecologists may be interested in any difference, whatever its sign, or only in the cases where O_{ij} is significantly higher than E_{ij} (preference) or significantly lower (avoidance, exclusion).

Bishop *et al.* (1975: 136 *et seq.*) describe three statistics for measuring the difference between O and E. They can be used for two-way or multiway contingency tables. The three statistics are the components of X^2_P, the components of X^2_W, and the Freeman-Tukey deviates:

$$\text{component of } X^2_P\text{: } (O - E) / \sqrt{E} \qquad \textbf{(6.26)}$$

$$\text{component of } X^2_W\text{: } 2\, O \log_e(O/E) \qquad \textbf{(6.27)}$$

$$\text{Freeman-Tukey deviate: } \sqrt{O} + \sqrt{O+1} - \sqrt{4E+1} \qquad \textbf{(6.28)}$$

These statistics are available in various computer packages. A critical value has been proposed by Bishop *et al.* (1975) for testing the significance of statistics 6.26 and 6.28:

$$\sqrt{\chi^2_{[\nu,\,\alpha]} / (\text{no.cells})}$$

E_{ij} is said to be significantly different from O_{ij} when the absolute value of the statistic, for cell (i, j), is larger than the critical value. According to Sokal & Rohlf (1995), however, the above critical value often results in a type I error much greater than the nominal α level. These authors use instead the following approximate criterion to test Freeman-Tukey deviates:

$$\sqrt{\nu\ \chi^2_{[1,\,\alpha]} / (\text{no.cells})} \qquad \textbf{(6.29)}$$

When the (absolute) value of the Freeman-Tukey deviate is larger than or equal to the criterion, one concludes that $E_{ij} \neq O_{ij}$ at significance level α for that cell. Authors often recommend to only test the cells where $5 \leq E_{ij} \leq (n - 5)$. Neu *et al.* (1974) recommended to apply a Bonferroni or Holm correction (Box 1.3) to significance level α in order to account for multiple testing. An example of this method, with Bonferroni correction for the number of tested cells, is presented in Table 6.7.

Test of standardized residuals

Alternatively, Haberman (1973) proposed a test of the components of X^2_P (eq. 6.26), which are also called *standardized residuals* and are represented by the symbol e_{ij}. The standard error of e_{ij} is the square root of the maximum likelihood estimate of its asymptotic variance:

$$var_{ij} = \left(1 - \frac{\text{row sum}_i}{n}\right)\left(1 - \frac{\text{column sum}_j}{n}\right)$$

Table 6.7 Test of Freeman-Tukey deviates (eq. 6.28) in individual cells of a contingency table. The observed and expected values are taken from Table 6.4. Only 8 of the 16 deviates are tested because the others, identified by an asterisk, had expected values smaller than 5 and could therefore not be tested. Absolute values larger than or equal to the criterion (eq. 6.29) with Bonferroni correction for 8 simultaneous tests, $[9\ \chi^2_{[1,\,0.05/8]}\ /\ 8]^{1/2} = [9 \times 7.48 / 8]^{1/2} = 2.90$, are in bold. These values identify the cells in which the number of observations (O_{ij}) significantly ($p < 0.05$) differs (higher or lower as shown by the sign) from the corresponding expected frequencies (E_{ij}). The overall null hypothesis (H_0: complete independence of descriptors **a** and **b**) had been rejected first (Table 6.4), before testing the significance of the observed values in individual cells of the table.

	b_1	b_2	b_3	b_4
a_1	**3.23**	–1.33	0.06	**–3.12**
a_2	**–4.57**	**3.49**	**–4.57**	0.91
a_3	–3.00 *	–3.00 *	–3.00 *	3.87 *
a_4	–3.00 *	–3.00 *	3.87 *	–3.00 *

* No test because $E_{ij} < 5$ (Table 6.4).

where n is the total number of observations in the contingency table. Dividing e_{ij} by $\sqrt{var_{ij}}$ produces an *adjusted residual* statistic Z_{ij}:

$$Z_{ij} = \frac{e_{ij}}{\sqrt{var_{ij}}} \qquad \textbf{(6.30)}$$

which is distributed like a standard normal deviate. That test is also described by Everitt (1977). When $|Z_{ij}|$ is larger than or equal to the critical value $z_{[1 - (\alpha\ /\ 2\ \text{no. tests})]}$ read from a table of standard normal deviates (z-table), one concludes that O_{ij} is significantly different from E_{ij} at significance level α. Division by the number of simultaneous tests is the Bonferroni correction (Box 1.3). Statistics higher than the critical value z are in bold-face type in Table 6.8. The conclusions drawn from Tables 6.7 and 6.8 may not be identical.

Comparing Table 6.4 to Tables 6.7 and 6.8 shows that considering only the cells where $O_{ij} > E_{ij}$ may lead to conclusions which, without necessarily being incorrect, are subject to some risk of error. Tables 6.7 and 6.8 show, for instance, that dominant species a_1 is significantly over-represented in environmental condition b_1 and under-represented in b_4, suggesting that b_1 is favourable whereas b_4 is adverse to the species.

Table 6.8 Test of standardized residuals using the *Z*-statistic (eq. 6.30). Only 8 of the 16 deviates are tested because the others, identified by an asterisk, had expected values smaller than 5 and could therefore not be tested. The observed and expected values are taken from Table 6.4. Absolute values of *Z* larger than or equal to the critical value $z_{[1-(0.05/2\times8)]} = z_{0.9969} = 2.73$ are in boldface; the correction is for 8 simultaneous tests. The bold values identify cells in which the number of observations (O_{ij}) significantly ($p < 0.05$) differs (higher or lower, as shown by the sign) from the corresponding expected frequency (E_{ij}).

	b_1	b_2	b_3	b_4
a_1	**6.32**	–2.11	0.00	**–4.22**
a_2	**–3.65**	**6.09**	**–3.65**	1.22
a_3	–2.39 *	–2.39 *	–2.39 *	7.17 *
a_4	–2.39 *	–2.39 *	7.17 *	–2.39 *

* No test because $E_{ij} < 5$ (Table 6.4).

Ecological application 6.4

Legendre *et al*. (1982) explored the relationship between the abundance of phytoplankton and vertical stability of the water column in a coastal embayment of Hudson Bay (Canadian Arctic). Surface waters are influenced by the plume of the nearby Great Whale River. There were intermittent phytoplankton blooms from mid-July through mid-September. In order to investigate the general relationship between phytoplankton concentrations (chlorophyll *a*) and the physical conditions, chl *a* and salinity data from 0 and 5 m depths were allocated to a contingency table (Table 6.9). The null hypothesis of independence being rejected, the correspondence between the two descriptors rests in four cells. (1) At high salinities (> 22), there is a significantly small number of high chl *a* observations and (2) a significantly high number of low chl *a* values. At intermediate salinities (18-22), (3) high chl *a* observations are significantly numerous, whereas (4) low chl *a* observations are significantly infrequent. At low salinities (< 18), the numbers observed are not significantly different from the frequencies expected under the null hypothesis of independence.

Table 6.9 shows that, on the one hand, high chl *a* concentrations were positively associated with intermediate salinities, whereas they were much reduced in waters of high salinity. On the other hand, low chl *a* concentrations were characteristically infrequent in waters of intermediate salinities and frequent at high salinities. The overall interpretation of these results, which also took into account estimates of the vertical stability of the water column (Richardson number), was as follows: (1) strong vertical mixing led to high salinities at the surface; this mixing favoured nutrient replenishment, but dispersed phytoplankton biomass over the water column; (2) low salinity conditions were not especially favourable nor adverse to phytoplankton,

Table 6.9 Contingency table: chlorophyll *a* concentrations as a function of salinity in the surface waters of Manitounuk Sound (Hudson Bay, Canadian Arctic). In each cell: observed (O_{ij}) and expected (E_{ij}, in parentheses) frequencies, and adjusted residual (Z, eq. 6.30) to test the hypothesis that $O_{ij} = E_{ij}$ ($\alpha = 0.05$) with correction for 5 simultaneous tests. Statistics in bold are larger than $z_{[1 - 0.05 / 2\times5]} = 2.58$, indicating that $O_{ij} \neq E_{ij}$. Total no. observations $n = 207$. $X^2_W = 33.47$ with Williams correction ($\nu = 2$, $p < 0.001$); hence the hypothesis of independence between chl *a* and salinity is rejected.

Chlorophyll *a* (mg m^{-3})	Salinity 6-18	18-22	22-26
	2	22	7
1.5-6.1	(3.29)	(8.09)	(19.62)
(high values)	–0.82 *	**6.17**	**–5.10**
	20	32	124
0-1.5	(18.71)	(45.91)	(111.38)
(low values)	0.82	**–6.17**	**5.10**

* Statistic not tested because $E_{ij} < 5$.

i.e. stratification was favourable, but dilution by water from the nearby river was adverse; (3) intermediate salinities were associated with intermittent conditions of stability; under such conditions, both the high nutrient concentrations and the stability of the water column were favourable to phytoplankton growth and accumulation. Intermittent summer blooms thus occurred upon stabilization of the water column, as a combined result of wind relaxation and fortnightly tides.

6.5 Species diversity

Biodiversity is a most important synthetic concept for ecology. It can be studied at all levels of organization of Life, from genes to ecosystems. Loreau (2010) gives a clear account of the importance of biodiversity science for both fundamental and applied ecology. He addresses, among other topics, the present crisis of diversity on Earth and the possibility of a sixth mass extinction, the socio-economic values of diversity within the context of ecological services, various frontiers of diversity science, the (controversial) linking of diversity science and policy, and finally, the need to build a new relationship between Humanity and Nature. The author also provides a well organised summary of different measures of diversity (see his Chapter 2). In the study of ecological communities, species diversity indices, discussed in the present section,

are synthetic biotic indices that capture multidimensional information relative to the species composition of an assemblage or a community.

Diversity is often called "biodiversity" nowadays. The addition of prefix "bio" before "diversity" has not changed the original concept or the way diversity is measured in ecology. Interested readers could look at the discussion of "diversity" versus "biodiversity" in Longhurst (2007, pp. 23-24).

The distribution of a quantitative variable is characterized by its *dispersion* around its mean, as shown in Sections 4.1 and 4.3. The parametric and nonparametric measures of dispersion are the *variance* (eq. 4.3) and the *range*, respectively. These two measures do not apply to qualitative variables, for which the *number of states* (q) may be used as a simple measure of dispersion. However, this measure does not take advantage of the frequency distribution of observations among the states, which is known in many instances. When the relative frequencies of the states are available, eq. 6.1 may be used to measure the dispersion of the qualitative variable:

$$H = -\sum_{i=1}^{q} \mathrm{p}_i \log \mathrm{p}_i$$

where p_i is the relative frequency or proportion (on a 0-1 scale) of observations in state (species) i. Species with frequency 0 disappear from the calculation because $\lim_{\mathrm{p} \to 0} (\mathrm{p} \log \mathrm{p}) = 0$. This formula can be rewritten as:

$$H = \frac{1}{n} \sum_{i=1}^{q} -(\log n_i - \log n)\, n_i$$

where n is the total number of organisms and n_i is the number of organisms belonging to species i. The latter equation is similar to the formula used to calculate the variance of n objects divided into q classes:

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{q} (y_i - \bar{y})\, f_i$$

where f_i is the frequency of the i-th class. In ecology, H is widely used to measure the *diversity* of a species assemblage; it is generally computed for each sampling site separately (alpha diversity; see Subsection 6.5.3). In species diversity studies, the qualitative descriptor is the list of the q species present and each state of that descriptor corresponds to a species name. Both the number of species q and entropy H belong to the same family of generalized entropies (eq. 6.31, below).

In assemblages of biological species, there are generally several species represented by a single or a few individuals, and a few species that are very abundant. The few abundant species often account for many more individuals than all the rare

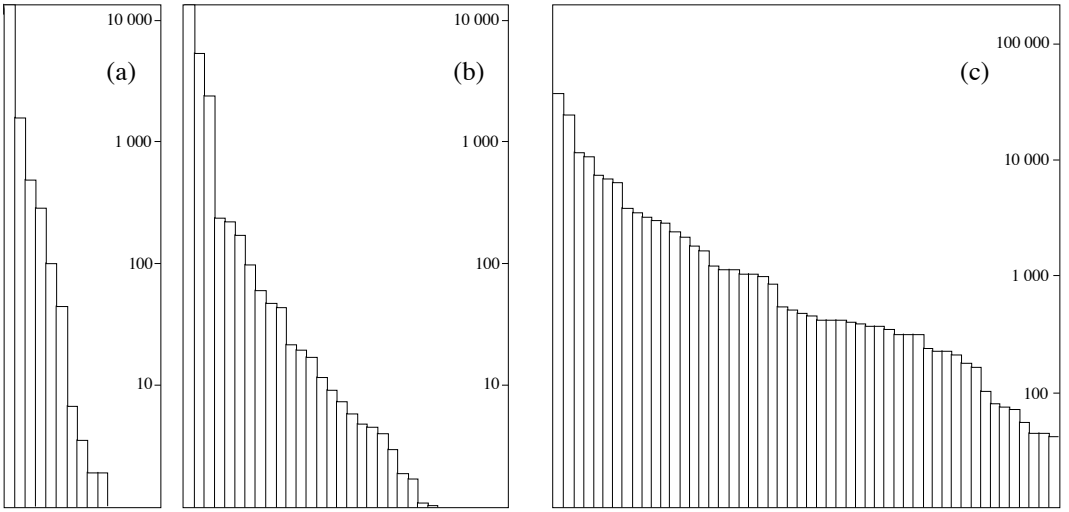


Figure 6.2 Fish catches (abundances) in (a) the Barents Sea, (b) the Indian Ocean, and (c) the Red Sea. Along the abscissa, species are arranged in order of decreasing frequencies. The histogram ordinates are logarithmic. Adapted from Margalef (1974).

species together. Figure 6.2 shows, in order of decreasing frequencies, the abundances of fish species caught in the Barents Sea, the Indian Ocean, and the Red Sea. The three water bodies clearly differ in both the number of species caught and the shape of their abundance distributions. Diversity indices must be applicable to any type of species assemblage regardless of the shape of the abundance distribution. One parameter of the distribution is clearly the *number of species*; another is the *shape of the distribution*. An alternative approach to describing a species frequency distribution with these two parameters is to combine them in a single index, e.g. the entropy measure H. Species diversity may thus be defined as a *measure of species composition, in terms of both the number of species and their relative abundances*.

Species diversity

It is generally not useful to measure species diversity of a whole community (e.g. primary, secondary, and tertiary producers and decomposers), because of the different roles played by various species in an ecosystem. It is better (Hurlbert, 1971; Pielou, 1975) to restrict the study of species diversity (and of the underlying theoretical phenomena, e.g. competition, succession) to a single *taxocene*. A taxocene is a set of species belonging to a given supraspecific taxon that make up a natural ecological community or, in other words, that represent a taxonomic segment of a community or association (Chodorowski, 1959; Hurlbert, 1971; Whittaker, 1972). The supraspecific taxon must be such that its member species are about the same size, have similar life histories, and compete over both ecological and evolutionary time for a finite amount of similar resources (Deevey, 1969). A taxocene occupies a limited

Taxocene

segment in space and in the environment. For these reasons, the following information about the reference population should accompany any measure of diversity: (1) the spatial boundaries of the region or volume within which the population is found and a description of the sampling method; (2) the temporal limits within which the observations have been made; (3) the taxocene under study (Hurlbert, 1971; Pielou, 1975).

Sampling sites may harbour species that differ much in size or role in the environment. This may occur, for example, when all plants in quadrats (ligneous and herbaceous) are counted, or when species at different developmental stages are collected (e.g. counting saplings as equivalent to adult trees). Comparisons of diversity indices with production or environmental variables may be easier in such cases if species diversity is computed, not from numbers of individuals, but instead from measures of biomass (Wilhm, 1968) or dry mass, productivity (Dickman, 1968), fecundity, or any other appropriate measure of energy transfer.

Species diversity indices may be used to compare successive observations from the same community (time series: O mode, Fig. 7.1) or sampling sites from different areas (Q mode). Coefficients in Chapter 7 compare objects by combining paired information available for each species. In contrast, diversity indices pool the multispecies information into a single value for each sampling unit, before comparing them.

Over the years, several formulae have been proposed in the ecological literature for measuring species diversity. The present section only describes the main indices that are found in the modern literature. Species diversity has been the subject of detailed discussions. Early reviews were presented in the milestone books of Pielou (1969, 1975) and Margalef (1974) and in the review paper of Peet (1974). A recent account linking species diversity to ecological theory is found in Loreau (2010).

1 — Diversity

Hill (1973a) and Pielou (1975) noted that the three diversity indices mostly used by ecologists are specific cases of the *generalized entropy* formula of Rényi (1961):

$$H_a = \frac{1}{1-a} \log \sum_{i=1}^{q} p_i^a \qquad \textbf{(6.31)}$$

where a is the order of the entropy measure ($a = 0, 1, 2, \ldots$), q is the number of species, and p_i is the relative frequency or proportion of species i. This formula gives an indeterminate result for $a = 1$. One can show, however, that the limit of this equation when a tends towards 1 from below (i.e. from 0 to 1) or from above (i.e. from 2 to 1) is the Shannon entropy formula, eqs. 6.1 and 6.34a.

Hill (1973a) prefers the corresponding diversity numbers:

Diversity number

$$N_a = \exp H_a \qquad \textbf{(6.32)}$$

The first three Rényi entropies H_a (of orders $a = 0$ to 2) and the corresponding Hill diversity numbers N_a are:

$$\text{(a) } H_0 = \log q \qquad \text{(b) } N_0 = q \qquad \textbf{(6.33)}$$

$$\text{(a) } H_1 = -\sum \mathrm{p}_i \log \mathrm{p}_i = H \qquad \text{(b) } N_1 = \exp H_1 \qquad \textbf{(6.34)}$$

$$\text{(a) } H_2 = -\log \sum \mathrm{p}_i^2 = -\log(\lambda) \qquad \text{(b) } N_2 = \lambda^{-1} \qquad \textbf{(6.35)}$$

Hill (1973a) noted that increasing the order a diminishes the relative weights of rare species in the resulting index: when $a = 0$, the data are transformed to presence-absence form where rare and abundant species have the same importance. In a review of the topic, Peet (1974) proposed other ways of creating families of diversity indices. Let us examine the first three orders of eq. 6.31 in more detail.

Number of species

1. *Entropy of order a = 0.* — The *number of species q* (eq. 6.33b) is the index of diversity most often used in ecology. It goes back to Patrick (1949):

$$\textit{Diversity} = q \qquad \textbf{(6.36)}$$

It is more affected by the presence of rare species than higher-order indices. The number of species can also be seen as a component of other diversity indices (e.g. J, eq. 6.45 in Subsection 6.5.2).

As the size of the sampling units increases, additional rare species appear. This is a problem with all diversity indices and it is at its worst in eq. 6.36. It is incorrect to compare the diversities of sampling units that have different sizes because diversity measures are not additive (Subsection 1.4.2). This point has been empirically shown by He *et al*. (1996). The problem can be resolved by calculating the numbers of species that sampling units would contain if they all had the same size. This can be done using Sanders' (1968) *rarefaction method*, whose formula was corrected by Hurlbert (1971). In this method, a constant number of organisms is used to make the sampling units comparable, instead of the physical size of sampling unit in m^2 or litre. The formula computes the expected number of species q' in a standardized sampling unit of n'

organisms, for example 1000 organisms, from a nonstandard sampling unit containing q species, a total of n organisms, and n_i organisms belonging to each species i:

$$E(q') = \sum_{i=1}^{q} \left[1 - \frac{\binom{n - n_i}{n'}}{\binom{n}{n'}} \right] \qquad \textbf{(6.37)}$$

where $n' \leq (n - n_1)$, n_1 being the number of individuals in the most abundant species (y_1), and the terms in parentheses are combinations. For example:

$$\binom{n}{n'} = \frac{n!}{n'!\,(n-n')!}$$

Shannon's entropy

2. *Entropy of order a = 1.* — Margalef (1958) proposed to use Shannon's entropy H (eqs. 6.1 and 6.34a) as an index of species diversity:

$$H = -\sum_{i=1}^{q} p_i \log p_i$$

The properties of H as a measure of diversity are the following:

• $H = 0$ (minimum value), when the sampling unit contains a single species. H increases with the number of species.

• For a given number of species, H is maximum when the organisms are equally distributed among the q species: $H = \log q$. For a given number of species, H is lower when there is stronger dominance in the sampling unit by one or a few species (e.g. Figs. 6.1a and b). The actual value of H depends on the base of logarithms (2, e, 10, or other). This base must always be reported since it sets the scale for the H values.

• Like the variance, diversity can be partitioned into different components. In particular, the calculation of diversity can take into account not only the proportions of the different species but also those of genera, families, etc. When diversity is partitioned into a component for *genera* and a component for *species within genera*, two adaptive levels can be explored among the environmental descriptors, and diversity H can be partitioned using eqs. 6.10-6.12. The total diversity, $H = A + B + C$, which is calculated using the proportions of species without taking into account those of genera, is equal to the diversity with respect to genera, $H(G) = A + B$, plus that of species within genera, $H(S \mid G) = C$. The latter is calculated as the sum of the diversities H within genera, weighted by the proportions of individuals in the genera. The formula is:

$$H = H(G) + H(S \mid G) \qquad \textbf{(6.38)}$$

This same calculation may be extended to other systematic categories. Considering, for example, the categories family (F), genus (G), and species (S), diversity can be partitioned into the following hierarchical components:

$$H = H(F) + H(G \mid F) + H(S \mid G,F) \qquad \textbf{(6.39)}$$

Using this approach, Lloyd *et al.* (1968) measured hierarchical components of diversity for communities of reptiles and amphibians in Borneo.

Most diversity indices share the first two properties above, but only the indices derived from eq. 6.31 have the third one (Daget, 1980). The probabilistic interpretation of H refers to the *uncertainty* about the identity of an organism chosen at random in a sampling unit. The uncertainty is small when the sampling unit is dominated by a few species or when the number of species is small. These two situations correspond to low H values.

In principle, H should only be used when a sample is drawn from a theoretically infinite population, or at least a population large enough that sampling does not modify it in a noticeable way. In cases of samples drawn from small populations, or samples whose representativeness is unknown, it is theoretically better, according to Pielou (1966), to use Brillouin's formula (1956), proposed by Margalef (1958) for computing diversity H. This formula was introduced in Section 6.1 to calculate the information per symbol in a message (eq. 6.3):

$$H = (1/n) \log[n! / (n_1!\, n_2! \ldots n_i! \ldots n_q!)]$$

where n_i is the number of individuals in species i and n is the total number of individuals in the collection. Brillouin's H corresponds to sampling *without* replacement (and is thus more exact) whereas Shannon's H applies to sampling *with* replacement. In practice, H computed with either formula is the same to several decimal places, unless samples are so small that they should not be used to estimate species diversity in any case. Species diversity cannot, however, be computed on measures of biomass or energy transfer using Brillouin's formula.

3. *Entropy of order a = 2.* — Simpson (1949) proposed an index of species diversity based on the probability that two interacting individuals of a population belong to the same species. This index is frequently used in ecology. When randomly drawing, without replacement, two organisms from a sampling unit containing q species and n individuals, the probability that the first organism belong to species i is n_i/n and that the second also belong to species i is $(n_i - 1)/(n - 1)$. The combined probability of the two events is the product of their separate probabilities. Simpson's

Concentration

concentration index (λ) is the probability that two randomly chosen organisms belong to the same species, i.e. the sum of the combined probabilities for the different species:

$$\lambda = \sum_{i=1}^{q} \frac{n_i}{n} \frac{(n_i - 1)}{(n-1)} = \frac{\sum_{i=1}^{q} n_i (n_i - 1)}{n(n-1)}$$

When n is large, n_i is almost equal to $(n_i - 1)$, so that the above equation becomes:

$$\lambda = \sum_{i=1}^{q} \left(\frac{n_i}{n}\right)^2 = \sum_{i=1}^{q} p_i^2 \qquad \textbf{(6.40)}$$

which corresponds to the summation in eq. 6.35a. This index may be computed from numbers of individuals, or from measures of biomass or energy transfer. The higher is the probability that two organisms be conspecific, the smaller is the diversity of the sampling unit. For this reason, Greenberg (1956) proposed to measure species diversity as:

$$\textit{Diversity} = 1 - \lambda \qquad \textbf{(6.41)}$$

which is also the *probability of interspecific encounter* (Hurlbert, 1971). Pielou (1969) showed that this index is an unbiased estimator of the diversity of the population from which the sample has been drawn. This index is also known in ecology as the *Gini coefficient*, because it was originally proposed by economist Corrado Gini (1912) as an index of "mutability" or diversity. In the same 1912 paper, Gini also defined an index of inequality, which is widely used in economics under the name of … *Gini coefficient*. Hence the Gini coefficient of ecologists is not the same as that of economists.

Because eq. 6.41 is more sensitive than H to changes in the abundances of the few very abundant species, Hill (1973a) recommended to use instead:

$$\textit{Diversity} = \lambda^{-1} \qquad \textbf{(6.42)}$$

which is diversity number N_2 of eq. 6.35b. Hill (1973a) also showed that this index is linearly related to $\exp H$ (eq. 6.34b).

Margalef & Gutiérrez (1983) proposed the following expression, which combines eqs. 6.41 and 6.42:

$$\textit{Diversity} = \frac{1-\lambda}{\lambda} = \sum_{i \neq j} p_i p_j \Big/ \sum_{i=1}^{q} p_i^2 \qquad \textbf{(6.43)}$$

Note that each pair (i, j), for $i \neq j$, is counted twice in the expression $\sum p_i p_j$. This diversity formula is the ratio of the probability that two individuals taken at random

belong to different species, to the probability that they pertain to the same species. It is the maximum number of interspecific interactions normalized by the maximum number of intraspecific interactions.

Biodiversity indices that integrate phylogenetic information have been proposed by Helmus *et al*. (2007).

Functional diversity

Functional diversity refers to the diversity of ecological processes that maintain interactions among the components of an ecosystem. It is estimated through the diversity of species traits and functions in a study area. Several functional diversity indices have been proposed by Rao (1982), Petchey & Gaston (2002), Botta-Dukát (2005), Villéger *et al*. (2008), Laliberté & Legendre (2010), and others. In practice, these indices are computed from species functional traits (quantitative or qualitative variables) weighted by species abundances.

2 — Evenness, equitability

Several authors, for example Margalef (1974), prefer to directly interpret species diversity as a function of physical, geographical, biological, or temporal variables, whereas others consider that species diversity consists of two components, which should be interpreted separately. These two components are the *number of species* and the *evenness* of their frequency distribution. Although the concept of evenness had been introduced by Margalef (1958), it was formally proposed by Lloyd & Ghelardi (1964) for characterizing the *shape* of distributions such as those in Fig. 6.2, where the component "number of species" corresponds to the length of the abscissa. In the literature "evenness" and "equitability" are synonyms terms (Lloyd & Ghelardi, 1964; see also the review of Peet, 1974). Several indices of evenness have been proposed.

1. The simplest approach to evenness consists in comparing the measured diversity to the corresponding maximum value. When using H (eqs. 6.1 and 6.34a), diversity takes its maximum value when all species are equally represented. In such a case,

$$H_{\max} = -\sum_{i=1}^{q} \frac{1}{q} \log \frac{1}{q} = \log q \qquad \textbf{(6.44)}$$

Pielou's evenness

where q *is* the number of species. Evenness (J) is computed as (Pielou, 1966):

$$J = H / H_{\max} = \left(-\sum_{i=1}^{q} p_i \log p_i \right) / \log q \qquad \textbf{(6.45)}$$

which is a ratio, whose value is independent of the base of logarithms used for the calculation. Using the terms defined by Hill (1973a, eqs. 6.31-6.35), Daget (1980) rewrote eq. 6.45 as the ratio of entropies of orders 1 (eq. 6.34a) and 0 (eq. 6.33a):

$$J = H_1 / H_0 \qquad \textbf{(6.46)}$$

Equations 6.44 and 6.45 show that diversity H combines the *number of species* (q) and the *evenness* of their distribution (J):

$$H = JH_{max} = J\log q \quad \textbf{(6.47)}$$

Hurlbert's evenness

2. Hurlbert (1971) proposed an evenness index based on the minimum and maximum values of diversity. Diversity is minimum when one species is represented by $(n - q + 1)$ organisms and the $(q - 1)$ others by a single organism. According to Hurlbert, the following indices are independent of q:

$$J = (D - D_{min}) / (D_{max} - D_{min}) \quad \textbf{(6.48)}$$

$$1 - J = (D_{max} - D) / (D_{max} - D_{min}) \quad \textbf{(6.49)}$$

Patten's redundancy

Equation 6.48 was proposed by Patten (1962) as a measure of *redundancy* (see Section 6.1). The two indices can be computed for any diversity index D.

Broken stick model

3. Instead of dividing the observed diversity by its maximum value, Lloyd & Ghelardi (1964) proposed to use a model based on the *broken stick distribution* (Barton & David, 1956; MacArthur, 1957). To generate this distribution, a set of individuals is taken as equivalent to a stick of unit length which is broken randomly into a number of pieces (i.e. in the present case, the number of species q). The divisor in the evenness formula is the diversity computed from the lengths of the pieces of the randomly broken stick. The expected lengths (E) of the pieces of the broken stick (species) y_i are given, in decreasing order, by the successive terms of the following series (Pielou, 1975), corresponding to the successive values $i = 1, 2, \ldots, q$, for a given number of species q:

$$E(y_i) = q^{-1} \sum_{x=i}^{q} x^{-1} \quad \textbf{(6.50)}$$

For example, for $q = 3$ species, eq. 6.50 gives the following lengths for species $i = 1$ to 3: 0.6111, 0.2778, and 0.1111, respectively (R function: Section 6.6). Diversity of this series is computed using the formula for H (eq. 6.1 or 6.34a):

$$M = -\sum_{i=1}^{q} E(y_i) \log E(y_i) \quad \textbf{(6.51)}$$

The evenness index of Lloyd & Ghelardi (1964), which they called *equitability*, is similar to eq. 6.45, with M being used instead of H_{max}:

$$J = H / M \quad \textbf{(6.52)}$$

In the same paper, Lloyd & Ghelardi proposed another evenness index:

$$J = q' / q \qquad \textbf{(6.53)}$$

where q is the observed number of species and q' is the number of species for which the broken stick model predicts the observed diversity H, i.e. $H(q) = M(q')$. Values computed with eq. 6.52 or 6.53 are usually, but not always, smaller than one. Indeed, it happens that biological populations are more diversified than predicted by the broken stick model.

Functional evenness

4. Troussellier & Legendre (1981) described an *index of functional evenness*, for studying bacterial assemblages. In such assemblages, the species level is often poorly defined. The index bypasses the step of species identification, using instead as data the set of binary biochemical (and other) descriptors that characterize the microbial isolates. The authors showed that their index has the usual properties of an evenness measure. Functional evenness J of a bacterial sampling unit is defined as:

$$J = \frac{I}{I_{\max}} = \frac{1}{c \log 0.5} \sum_{i=1}^{c} [\, p_i \log p_i + (1 - p_i) \log (1 - p_i) \,] \qquad \textbf{(6.54)}$$

where I and $I_{\max}$ are measures of information, c is the number of binary descriptors used, and p_i is the proportion of positive responses to test i.

Evenness indices 6.44, 6.47, 6.51, and 6.52 all suffer from the problem that they depend on field estimation of the number of species in the population; in other words, q is not a fixed and known value but a random variable. Because the true value of q is not known and cannot be estimated from the data, there is no formula for computing a standard error (and, thus, a confidence interval) for these estimates of J. This point has been stressed by Pielou (1975) for eq. 6.45. This is not the case with eq. 6.54, where the denominator of J is a constant ($I_{\max} = c \log 0.5$ where c is the number of binary descriptors used in the calculation). Several methods may be used for computing the confidence interval of J (e.g. the jackknife, briefly described at the end of Subsection 1.2.4). Legendre *et al.* (1984b) provided examples where the computation of confidence intervals for J, measured during biodegradation experiments, showed that significant changes had taken place, at some point in time, in the structure of the bacterial assemblages involved in the biodegradation processes.

In varying environments, the ecological interpretation of the two components of diversity (eq. 6.47) could be carried out, for example, along the lines proposed by Legendre (1973). (1) The *number of species* may be a function of the stability of the environment. Indeed, a more stable environment entails a higher degree of organization and complexity of the food web (Margalef, 1958), so that such an environment contains more niches and, thus, more species. The number of species is proportional to the number of niches since, by definition, the realized niche of a species is the set of environmental conditions that this species does not share with any

other sympatric species (Hutchinson, 1957, 1965). This approach has the advantage of linking species diversity to environmental diversity. (2) The *evenness of species distribution* may be inversely related to the overall biological activity in the studied environment; the lower the evenness, the higher the biological activity (e.g. production, life cycles, energy flows among trophic levels). On a seasonal basis, another factor may contribute to lower the evenness. In an environment where interspecific competition is low (high evenness), seasonal reduction of resources or deterioration of weather conditions could induce stronger competition and thus favour some species over others, which would decrease the evenness. The same is often observed in cases of pollution.

3— Species diversity through space

A most interesting property of species diversity is its organization through space. This phenomenon, which is now well known to community ecologists, was first discussed by Whittaker in two seminal papers (1960, 1972) where he described the alpha, beta and gamma diversity levels. The development of multiscale spatial analysis of communities (Chapter 14) is grounded in Whittaker's concept of beta diversity.

Alpha diversity

Alpha (α) diversity is the diversity in species composition at individual sites i (e.g. plots, quadrats; α_i in Fig. 6.3). The indices used for alpha diversity estimate, in different ways, the variance in the species identity of individuals observed at a given site. A monoculture, for example, has the lowest possible alpha diversity because there is no variance in species identity among the individuals. Alpha diversity is measured by one of Rényi's entropy indices H_0 (eq. 6.33a), H_1 (eq. 6.34a) or H_2 (eq. 6.35a), by Hill's diversity numbers N_0 (richness, eq. 6.33b), N_1 (eq. 6.34b) or N_2 (eq. 6.35b), or by some other indices such as Fisher's α logarithmic series parameter (Fisher *et al.*, 1943). The most commonly used indices are N_0, H_1 and N_2, mentioned in Fig. 6.3.

Gamma diversity

Gamma (γ) diversity is the diversity of the whole region of interest in a study (γ in Fig. 6.3). It is usually measured by pooling the observations from a group of sampling units (which form a *sample* in the statistical sense), i.e. a large number of sites from the area of interest, except in cases where the community composition of an entire area is known, e.g. the CTFS permanent forest plots[*]. Gamma diversity is measured using the same indices as alpha diversity.

Beta diversity

Beta (β) diversity is of different nature: it is conceptually the variation in species composition among sites in the geographic area of interest (Legendre *et al.,* 2005, Anderson *et al.,* 2006; β in Fig. 6.3). Its value will vary with the extent of the area, the physical size of the sampling units and the sampling interval in the area under study, which form three aspects of the study scale (Section 13.0). Studies of beta diversity can actually focus on two aspects of community structure (Anderson *et al.*, 2011). The

[*] A map of the *Center for Tropical Forest Science* (CTFS) forest plots, and details about each plot, are available on the Web page http://www.ctfs.si.edu/.

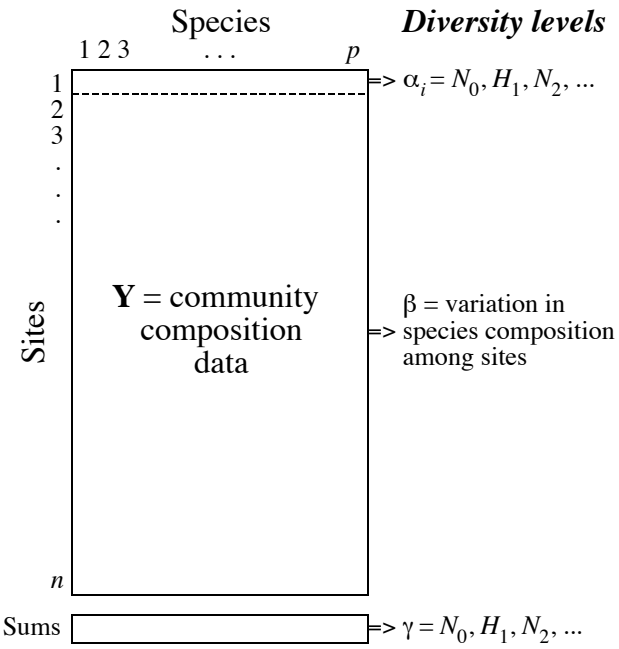


Figure 6.3 Species diversity indices are computed from the community composition data (matrix **Y**). Alpha (α) diversity indices are computed for individual sites (rows) *i*. Gamma diversity (γ) is computed from the vector of column sums of the data matrix using the same indices as for alpha diversity. Beta (β) diversity is of a different nature: it is the variation in community composition among sites. It cannot be computed with the usual entropy of diversity number indices.

first one is *turnover*, or the change in community composition between adjacent sampling units, explored by sampling along a spatial, temporal, or environmental gradient. The second is a *non-directional* approach to the study of *community variation* through space [or time]; it does not refer to any specific gradient but centres on the variation in community composition among the study units. The present section focuses on the second approach, as it links the concept of beta diversity with the analysis of the variation of community data matrices performed by the methods described in the following chapters.

If the variation in community composition is random and accompanied by biotic processes (e.g. reproduction) that generate spatial autocorrelation in the species data due to their limited dispersal (Subsection 1.1.1, model 2; Fig. 1.5, case 3), a gradient in species composition may appear (called a "false gradient" in Subsection 13.1.2) if the sampling area is small compared to the dispersal distance. Beta diversity can then be interpreted in terms of the rate of change, or *turnover*, in species composition along that gradient. Ecologists often refer to this turnover to explain beta diversity. The

community spatial structure is often more complex than a single gradient, however: if differentiation among sites is due to environmental factors, which may combine gradient-like and patchy geographic distributions, beta diversity should be analysed with respect to the hypothesized forcing environmental variables (Subsection 1.1.1, model 1; Fig. 1.5, case 4). In ecosystems, beta diversity may be caused concurrently by varying proportions of these two processes (i.e. induced spatial dependence and true autocorrelation due to biotic processes: Fig. 1.5, case 5). Chapter 14 will show how these two types of hypotheses about the processes that generate beta diversity can be disentangled.

Whittaker (1960, 1972) showed that beta diversity could be estimated using either presence-absence or quantitative species data. Ecologists use both types of measures to study beta diversity, although some researchers only refer to presence-absence data when they talk about the rate of species replacement, or turnover, along an ecological gradient. In the ordination literature, however, ecologists most often use species abundance data to study turnover rates by reference to the appearance and disappearance of species with unimodal distributions along gradients.

A first method, proposed by Whittaker (1960, 1972), for obtaining a global measure of beta diversity from species presence-absence data, is to compute the ratio of two diversity indices: $\beta = S/\bar{\alpha}$, where S is the number of species in a composite community composition vector representing the area of interest, and $\bar{\alpha}$ is the mean number of species observed at the sites that were used to compute S. This is a multiplicative approach, where S represents gamma diversity. The ratio $S/\bar{\alpha}$ indicates how many more species are present in the whole region than at an average site, and uses that value as the measure of beta diversity. Other beta diversity indices have been reviewed by Koleff *et al*. (2003), Magurran (2004), Tuomisto (2010) and Anderson *et al*. (2011).

An alternative, additive approach had been present in the literature since MacArthur *et al*. (1966), Levins (1968) and Allan (1975). It was revived by Lande (1996) and has been widely used since then (Veech *et al*., 2002). In that approach, $D_T = D_{among} - \bar{D}_{within}$ where D_T is the total (gamma) diversity. This approach can be applied to species richness N_0 (eq. 6.33b), Shannon information H_1 (eqs. 6.1 and 6.34a), or Simpson diversity $D = (1 - \lambda)$ (eq. 6.41); see Lande (1996) for details. Because diversities are variances, one recognizes an analysis of variance approach in that equation.

Whittaker (1960, 1972) suggested that beta diversity could also be estimated from distance matrices computed among sites. This approach is based on the fact that a distance between two sites, computed from community composition data, provides a measure of the variation, or beta diversity between these sites. Distance matrices computed using appropriate indices (Chapter 7) thus assess the pairwise beta diversity among all pairs of sites. To obtain an overall index of beta diversity over a group of sites, Whittaker (1972) suggested to use *the mean* (not the variance) of the distances among sites: "the mean CC [i.e. the distance coefficient that is the complement of

Jaccard's coefficient of community, $D = 1 - S_7$ in Table 7.2] for samples of a set compared with one another [...] is one expression [of] their relative dissimilarity, or beta differentiation" (Whittaker, 1972: 233). Whittaker derived his concept from the *index of biotal dispersity* suggested fifteen years before by Koch (1957). Whittaker thus acknowledged the fact that dissimilarities (i.e. distances, Chapter 7) are themselves measures of the differentiation between sites.

Total variation of **Y**

Box 6.1 shows that the total variation of a data matrix **Y**, e.g. the one shown in Fig. 6.3, can be computed either from **Y** itself or from a distance matrix **D** derived from **Y**. This equality is pertinent here as it shows the equivalence of Whittaker's overall measure of beta diversity computed from a distance matrix, **D**, and beta diversity defined as the variation in species composition among sites, which can be measured by the total variation in matrix **Y**, SS(**Y**). Indeed, SS(**Y**) can be computed from matrix **D** using eq. 6.56. For distance matrices that are not Euclidean but whose square root is Euclidean, one may use eq. 6.58. The distance functions that Whittaker (1972) was citing, i.e. $1 - S_7$ (Jaccard), $1 - S_8$ (Sørensen), D_9 (Whittaker), and D_{14} (percentage difference), pertain to that group. Box 6.1 shows that eq. 6.58 is a logical choice for the computation of SS(**Y**) for such distance functions.

To sum up, beta diversity can be estimated as the total variation in **Y** using two different equations: by computing eq. 6.55 from the raw data table **Y**, or computing eq. 6.56 on distances that have the Euclidean property, e.g. the Euclidean, chord, chi-square and Hellinger distances. Equation 6.58 is an alternative reasonable choice for distances whose square root is Euclidean, e.g. the (1 – Jaccard), (1 – Sørensen), Whittaker, and percentage difference distances.

An interesting observation is that for the chord and Hellinger distances, the maximum possible value of total variance Var(**Y**), computed by applying eq. 6.56 followed by eq. 6.57, is 1. The maximum values are obtained when all sites in table **Y** have entirely different species compositions. Similarly for community composition data transformed using the chord or Hellinger transformations (Section 7.7), the maximum possible value of Var(**Y**), computed using eq. 6.55 followed by eq. 6.57, is 1. Hence, using these transformations or distances, the estimates of beta diversity provided by Var(**Y**) are easily comparable since they fall in the range 0 to 1.

SS(Y), Var(Y) — Box 6.1

The total variation in a data matrix **Y** with n rows and p columns can be computed in two ways, which produce the same result.

• First method — Centre each column of **Y** on its mean using eq. 1.9 to obtain matrix $\mathbf{Y}_{\text{cent}} = [\, y_{\text{cent}.ij} \,]$, then compute the sum of these centred values squared:

$$\text{SS}(\mathbf{Y}) = \sum_{j=1}^{p} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \sum_{j=1}^{p} \sum_{i=1}^{n} y_{\text{cent}.ij}^{\,2} \qquad \textbf{(6.55)}$$

This is the total variation, or total sum of squares, of matrix **Y**. It is noted SS(**Y**), or e_k^2 in eq. 8.5.

• Second method — Compute a Euclidean distance matrix $\mathbf{D} = [D_{ih}]$ among the n rows of **Y** using distance function D_1 (eq. 7.32, Chapter 7). Then, calculate

$$\text{SS}(\mathbf{Y}) = \left(\sum_{i \neq h} D_{ih}^2 \right) / n \qquad \textbf{(6.56)}$$

using the $n(n-1)/2$ distances from the upper [or lower] triangular portion of **D**. SS(**Y**) computed in this way is called e_k^2 in eq. 8.6. The equivalence of these two ways of computing SS(**Y**) (Fig. 8.18) is demonstrated in Appendix 1 of Legendre & Fortin (2010).

The total variance in **Y** can be calculated from SS(**Y**) computed either way:

$$\text{Var}(\mathbf{Y}) = \text{SS}(\mathbf{Y})/(n-1) \qquad \textbf{(6.57)}$$

Besides eq. 6.56, there are three other ways of computing SS(**Y**) from **D**:

• SS(**Y**) is the trace of the Gower-centred distance matrix Δ_1 derived from **D** (eqs. 9.40 and 9.41, Chapter 9).

• SS(**Y**) is the sum of the eigenvalues of Δ_1, i.e. the eigenvalues of the principal coordinate analysis (PCoA) of **D**.

• SS(**Y**) is the total sum of squares of the principal coordinates of **D** (e.g. Table 9.9).

Box 6.1 (*continued*)

Equation 6.56 can be applied to any distance matrix **D**, Euclidean or not.

• Euclidean distances — For distances that have the Euclidean property (Tables 7.2 and 7.3), the rectangular matrix **Y**' obtained by principal coordinate analysis of **D** contains real numbers only. The distances among the rows of **Y**' computed using the Euclidean distance function D_1 (eq. 7.32) are equal to the distances in **D** (Subsection 9.3.3). Thus the total sum of squares in **Y**' computed with eq. 6.55 is equal to SS(**Y**) computed by applying eq. 6.56 to **D**.

Four of the Euclidean distance functions recommended for community composition data in Table 7.4 — the chord distance D_3 (eq. 7.35), the distance between species profiles D_{18} (eq. 7.53), the chi-square distance D_{16}, (eq. 7.55) and the Hellinger distance D_{17} (eq. 7.56) — have an additional property: eq. 6.55 computed from community composition data transformed using the chord (eq. 7.67), profile (eq. 7.68), chi-square (eq. 7.70) or Hellinger transformations (eq. 7.69) produces values of SS(**Y**) identical to those computed using eq. 6.56 with the chord, species profiles, chi-square and Hellinger distance matrices.

• Non-Euclidean distances — Examples of distance functions described in Chapter 7 that do not have the Euclidean property in their basic form are the Jaccard distance ($1 - S_7$), the Sørensen distance ($1 - S_8$), the percentage difference distance ($D_{14} = 1 - S_{17}$; D_{14} is called the Bray-Curtis distance in some computer packages), and the Whittaker distance (D_9); they may produce negative eigenvalues in principal coordinate analysis (PCoA, Section 9.3). For these distances, one can still compute eq. 6.56, but the corresponding matrix **Y**' of principal coordinates contains both real and complex (imaginary) axes (Subsection 9.3.4). Equation 6.55 can still be computed for **Y**' (McArdle & Anderson, 2001) with the result that SS(**Y**') is equal to the total sum of squares computed from **D** using eq. 6.56.

Ecologists may not be comfortable, however, in considering a matrix **Y**' that contains complex axes as a fair representation of community composition data. Luckily, there is another way: matrix $\mathbf{D}' = [D_{ih}^{0.5}]$, which contains the square roots of the distances, is Euclidean for these (and some other) distance functions, as shown in Tables 7.2 and 7.3. Hence, SS(**Y**) computed by applying eq. 6.56 to **D**' is equal to the total variation (eq. 6.55) of the rectangular data matrix **Y**" obtained by principal coordinate analysis (PCoA) of **D**', and this time **Y**" only contains real axes. So, for these non-Euclidean distance functions, because D_{ih} is equal to $\sqrt{D_{ih}^2}$, an appropriate formula for computing SS(**Y**) for the original matrix $\mathbf{D} = [D_{ih}]$ is:

$$\mathrm{SS}(\mathbf{Y}) = \left(\sum_{i \neq h} D_{ih} \right) / n \qquad \textbf{(6.58)}$$

6.6 Software

Two-way and multiway contingency table analysis are available in most commercial statistical software. The R language also offers functions implementing the methods described in this chapter.

1. In R, the standard function to conduct the Pearson chi-square test on a contingency table crossing two qualitative variables is ***chisq.test()*** of STATS; parametric and permutation tests are available in that function. Package SURVEY contains several functions to construct contingency tables and perform chi-square tests of association for survey data. Using cross-classifying factors, functions ***table()*** and ***ftable()*** of BASE construct two-way or multi-way contingency tables crossing factor levels. Function ***table.cont()*** in ADE4 plots contingency table data into a graph.

Function ***mantelhaen.test()*** of STATS performs a Cochran-Mantel-Haenszel chi-square test of interaction between two factors in three-way contingency tables. Also in STATS, function ***loglin()*** fits log-linear models to multidimensional contingency tables*.

2. R functions for studying diversity are found in packages BIODIVERSITYR, VEGAN and PICANTE. Rarefaction curves are computed by VEGAN's function ***rarefy()***. In a phylogenetic context, ***specaccum.psr()*** in PICANTE computes a rarefaction curve for phylogenetic species richness.

Function ***dbFD()*** of package FD computes seven functional diversity indices. Among these, Rao's (1982) quadratic entropy is also computed by functions ***divc()*** of ADE4 and ***raoD()*** of PICANTE. The broken-stick distribution is computed by function ***bstick()*** in VEGAN.

* A tutorial is available at the Web address http://ww2.coastal.edu/kingw/statistics/R-tutorials/loglin.html.