Chapter 1

Introduction

Biologists and environmental scientists today must contend with the demands of keeping up with their primary field of specialization, and at the same time ensuring that their set of professional tools is current. Those tools may include topics as diverse as molecular genetics, sediment chemistry, and small-scale hydrodynamics, but one tool that is common and central to most of us is an understanding of experimental design and data analysis, and the decisions that we make as a result of our data analysis determine our future research directions or environmental management. With the advent of powerful desktop computers, we can now do complex analyses that in previous years were available only to those with an initiation into the wonders of early mainframe statistical programs, or computer programming languages, or those with the time for laborious hand calculations. In past years, those statistical tools determined the range of sampling programs and analyses that we were willing to attempt. Now that we can do much more complex analyses, we can examine data in more sophisticated ways. This power comes at a cost because we now collect data with complex underlying statistical models, and, therefore, we need to be familiar with the potential and limitations of a much greater range of statistical approaches.

With any field of science, there are particular approaches that are more common than others. Texts written for one field will not necessarily cover the most common needs of another field, and we felt that the needs of most common biologists and environmental scientists of our acquaintance were not covered by any one particular text.

A fundamental step in becoming familiar with data collection and analysis is to understand the philosophical viewpoint and basic tools that underlie what we do. We begin by describing our approach to scientific method. Because our aim is to cover some complex techniques, we do not describe introductory statistical methods in much detail. That task is a separate one, and has been done very well by a wide range of authors. We therefore provide only an overview or refresher of some basic philosophical and statistical concepts. We strongly urge you to read the first few chapters of a good introductory statistics or biostatistics book (you can't do much better than Sokal & Rohlf 1995) before working through this chapter.

1.1 Scientific method

An appreciation of the philosophical bases for the way we do our scientific research is an important prelude to the rest of this book (see Chalmers 1999, Gower 1997, O'Hear 1989). There are many valuable discussions of scientific philosophy from a biological context and we particularly recommend Ford (2000), James & McCulloch (1985), Loehle (1987) and Underwood (1990, 1991). Maxwell & Delaney (1990) provide an overview from a behavioral sciences viewpoint and the first two chapters of Hilborn & Mangel (1997) emphasize alternatives to the Popperian approach in situations where experimental tests of hypotheses are simply not possible.

Early attempts to develop a philosophy of scientific logic, mainly due to Francis Bacon and John Stuart Mill, were based around the principle of induction, whereby sufficient numbers of confirmatory observations and no contradictory observations allow us to conclude that a theory or law is true (Gower 1997). The logical problems with inductive reasoning are discussed in every text on the philosophy of science, in particular that no amount of confirmatory observations can ever prove a theory. An alternative approach, and also the most commonly used scientific method in modern biological sciences literature, employs deductive reasoning, the process of deriving explanations or predictions from laws or theories. Karl Popper (1968, 1969) formalized this as the hypothetico-deductive approach, based around the principle of falsificationism, the doctrine whereby theories (or hypotheses derived from them) are *disproved* because proof is logically impossible. An hypothesis is falsifiable if there exists a logically possible observation that is inconsistent with it. Note that in many scientific investigations, a description of pattern and inductive reasoning, to develop models and hypotheses (Mentis 1988), is followed by a deductive process in which we critically test our hypotheses.

Underwood (1990, 1991) outlined the steps involved in a falsificationist test. We will illustrate these steps with an example from the ecological literature, a study of bioluminescence in dinoflagellates by Abrahams & Townsend (1993).

1.1.1 Pattern description

The process starts with observation(s) of a pattern or departure from a pattern in nature. Underwood (1990) also called these puzzles or problems. The quantitative and robust description of patterns is, therefore, a crucial part of the scientific process and is sometimes termed an observational study (Manly 1992). While we strongly advocate experimental methods in biology, experimental tests of hypotheses derived from poorly collected and interpreted observational data will be of little use.

In our example, Abrahams & Townsend (1993) observed that dinoflagellates bioluminesce when the water they are in is disturbed. The next step is to explain these observations.

1.1.2 Models

The explanation of an observed pattern is referred to as a model or theory (Ford 2000), which is a series of statements (or formulae) that explains why the observations have occurred. Model development is also what Peters (1991) referred to as the synthetic or private phase of the scientific method, where the perceived problem interacts with insight, existing theory, belief and previous observations to produce a set of competing models. This phase is clearly inductive and involves developing theories from observations (Chalmers 1999), the exploratory process of hypothesis formulation.

James & McCulloch (1985), while emphasizing the importance of formulating models in science, distinguished different types of models. Verbal models are non-mathematical explanations of how nature works. Most biologists have some idea of how a process or system under investigation operates and this idea drives the investigation. It is often useful to formalize that idea as a conceptual verbal model, as this might identify important components of a system that need to be included in the model. Verbal models can be quantified in mathematical terms as either empiric models or theoretic models. These models usually relate a response or dependent variable to one or more predictor or independent variables. We can envisage from our biological understanding of a process that the response variable might depend on, or be affected by, the predictor variables.

Empiric models are mathematical descriptions of relationships resulting from processes rather than the processes themselves, e.g. equations describing the relationship between metabolism (response) and body mass (predictor) or species number (response) and island area (first predictor) and island age (second predictor). Empiric models are usually statistical models (Hilborn & Mangel 1997) and are used to describe a relationship between response and predictor variables. Much of this book is based on fitting statistical models to observed data.

Theoretic models, in contrast, are used to study processes, e.g. spatial variation in abundance of intertidal snails is caused by variations in settlement of larvae, or each outbreak of

Mediterranean fruit fly in California is caused by a new colonization event (Hilborn & Mangel 1997). In many cases, we will have a theoretic, or scientific, model that we can re-express as a statistical model. For example, island biogeography theory suggests that the number of species on an island is related to its area. We might express this scientific model as a linear statistical relationship between species number and island area and evaluate it based on data from a range of islands of different sizes. Both empirical and theoretic models can be used for prediction, although the generality of predictions will usually be greater for theoretic models.

The scientific model proposed to explain bioluminescence in dinoflagellates was the "burglar alarm model", whereby dinoflagellates bioluminesce to attract predators of copepods, which eat the dinoflagellates. The remaining steps in the process are designed to test or evaluate a particular model.

1.1.3 Hypotheses and tests

We can make a prediction or predictions deduced from our model or theory; these predictions are called research (or logical) hypotheses. If a particular model is correct, we would predict specific observations under a new set of circumstances. This is what Peters (1991) termed the analytic, public or Popperian phase of the scientific method, where we use critical or formal tests to evaluate models by falsifying hypotheses. Ford (2000) distinguished three meanings of the term "hypothesis". We will use it in Ford's (2000) sense of a statement that is tested by investigation, experimentally if possible, in contrast to a model or theory and also in contrast to a postulate, a new or unexplored idea.

One of the difficulties with this stage in the process is deciding which models (and subsequent hypotheses) should be given research priority. There will often be many competing models and, with limited budgets and time, the choice of which models to evaluate is an important one. Popper originally suggested that scientists should test those hypotheses that are most easily falsified by appropriate tests. Tests of theories or models using hypotheses with high empirical content and which make improbable predictions are what

Popper called severe tests, although that term has been redefined by Mayo (1996) as a test that is likely to reveal a specific error if it exists (e.g. decision errors in statistical hypothesis testing – see Chapter 3). Underwood (1990, 1991) argued that it is usually difficult to decide which hypotheses are most easily refuted and proposed that competing models are best separated when their hypotheses are the most distinctive, i.e. they predict very different results under similar conditions. There are other ways of deciding which hypothesis to test, more related to the sociology of science. Some hypotheses may be relatively trivial, or you may have a good idea what the results can be. Testing that hypothesis may be most likely to produce a statistically significant (see Chapter 3), and, unfortunately therefore, a publishable result. Alternatively, a hypothesis may be novel or require a complex mechanism that you think unlikely. That result might be more exciting to the general scientific community, and you might decide that, although the hypothesis is harder to test, you're willing to gamble on the fame, money, or personal satisfaction that would result from such a result.

Philosophers have long recognized that proof of a theory or its derived hypothesis is logically impossible, because all observations related to the hypothesis must be made. Chalmers (1999; see also Underwood 1991) provided the clever example of the long history of observations in Europe that swans were white. Only by observing all swans everywhere could we "prove" that all swans are white. The fact that a single observation contrary to the hypothesis could disprove it was clearly illustrated by the discovery of black swans in Australia.

The need for disproof dictates the next step in the process of a falsificationist test. We specify a null hypothesis that includes all possibilities except the prediction in the hypothesis. It is much simpler logically to disprove a null hypothesis. The null hypothesis in the dinoflagellate example was that bioluminesence by dinoflagellates would have no effect on, or would decrease, the mortality rate of copepods grazing on dinoflagellates. Note that this null hypothesis includes all possibilities except the one specified in the hypothesis.

So, the final phase in the process is the experimental test of the hypothesis. If the null hypothesis is rejected, the logical (or research) hypothesis, and therefore the model, is supported. The model should then be refined and improved, perhaps making it predict outcomes for different spatial or temporal scales, other species or other new situations. If the null hypothesis is not rejected, then it should be retained and the hypothesis, and the model from which it is derived, are incorrect. We then start the process again, although the statistical decision not to reject a null hypothesis is more problematic (Chapter 3).

The hypothesis in the study by Abrahams & Townsend (1993) was that bioluminesence would increase the mortality rate of copepods grazing on dinoflagellates. Abrahams & Townsend (1993) tested their hypothesis by comparing the mortality rate of copepods in jars containing bioluminescing dinoflagellates, copepods and one fish (copepod predator) with control jars containing non-bioluminescing dinoflagellates, copepods and one fish. The result was that the mortality rate of copepods was greater when feeding on bioluminescing dinoflagellates than when feeding on non-bioluminescing dinoflagellates. Therefore the null hypothesis was rejected and the logical hypothesis and burglar alarm model was supported.

1.1.4 Alternatives to falsification

While the Popperian philosophy of falsificationist tests has been very influential on the scientific method, especially in biology, at least two other viewpoints need to be considered. First, Thomas Kuhn (1970) argued that much of science is carried out within an accepted paradigm or framework in which scientists refine the theories but do not really challenge the paradigm. Falsified hypotheses do not usually result in rejection of the over-arching paradigm but simply its enhancement. This "normal science" is punctuated by occasional scientific revolutions that have as much to do with psychology and sociology as empirical information that is counter to the prevailing paradigm (O'Hear 1989). These scientific revolutions result in (and from) changes in methods, objectives and personnel (Ford 2000). Kuhn's arguments have been described as relativ-

istic because there are often no objective criteria by which existing paradigms and theories are toppled and replaced by alternatives.

Second, Imre Lakatos (1978) was not convinced that Popper's ideas of falsification and severe tests really reflected the practical application of science and that individual decisions about falsifying hypotheses were risky and arbitrary (Mayo 1996). Lakatos suggested we should develop scientific research programs that consist of two components: a "hard core" of theories that are rarely challenged and a protective belt of auxiliary theories that are often tested and replaced if alternatives are better at predicting outcomes (Mayo 1996). One of the contrasts between the ideas of Popper and Lakatos that is important from the statistical perspective is the latter's ability to deal with multiple competing hypotheses more elegantly than Popper's severe tests of individual hypotheses (Hilborn & Mangel 1997).

An important issue for the Popperian philosophy is corroboration. The falsificationist test makes it clear what to do when an hypothesis is rejected after a severe test but it is less clear what the next step should be when an hypothesis passes a severe test. Popper argued that a theory, and its derived hypothesis, that has passed repeated severe testing has been corroborated. However, because of his difficulties with inductive thinking, he viewed corroboration as simply a measure of the past performance of a model, rather an indication of how well it might predict in other circumstances (Mayo 1996, O'Hear 1989). This is frustrating because we clearly want to be able to use models that have passed testing to make predictions under new circumstances (Peters 1991). While detailed discussion of the problem of corroboration is beyond the scope of this book (see Mayo 1996), the issue suggests two further areas of debate. First, there appears to be a role for both induction and deduction in the scientific method, as both have obvious strengths and weaknesses and most biological research cannot help but use both in practice. Second, formal corroboration of hypotheses may require each to be allocated some measure of the probability that each is true or false, i.e. some measure of evidence in favor or against each hypothesis. This goes to the heart of

one of the most long-standing and vigorous debates in statistics, that between frequentists and Bayesians (Section 1.4 and Chapter 3).

Ford (2000) provides a provocative and thorough evaluation of the Kuhnian, Lakatosian and Popperian approaches to the scientific method, with examples from the ecological sciences.

1.1.5 Role of statistical analysis

The application of statistics is important throughout the process just described. First, the description and detection of patterns must be done in a rigorous manner. We want to be able to detect gradients in space and time and develop models that explain these patterns. We also want to be confident in our estimates of the parameters in these statistical models. Second, the design and analysis of experimental tests of hypotheses are crucial. It is important to remember at this stage that the research hypothesis (and its complement, the null hypothesis) derived from a model is not the same as the statistical hypothesis (James & McCulloch 1985); indeed, Underwood (1990) has pointed out the logical problems that arise when the research hypothesis is identical to the statistical hypothesis. Statistical hypotheses are framed in terms of population parameters and represent tests of the predictions of the research hypotheses (James & McCulloch 1985). We will discuss the process of testing statistical hypotheses in Chapter 3. Finally, we need to present our results, from both the descriptive sampling and from tests of hypotheses, in an informative and concise manner. This will include graphical methods, which can also be important for exploring data and checking assumptions of statistical procedures.

Because science is done by real people, there are aspects of human psychology that can influence the way science proceeds. Ford (2000) and Loehle (1987) have summarized many of these in an ecological context, including confirmation bias (the tendency for scientists to confirm their own theories or ignore contradictory evidence) and theory tenacity (a strong commitment to basic assumptions because of some emotional or personal investment in the underlying ideas). These psychological aspects can produce biases in a given discipline that have important implications for our subsequent discussions on research

design and data analysis. For example, there is a tendency in biology (and most sciences) to only publish positive (or statistically significant) results, raising issues about statistical hypothesis testing and meta-analysis (Chapter 3) and power of tests (Chapter 7). In addition, successful tests of hypotheses rely on well-designed experiments and we will consider issues such as confounding and replication in Chapter 7.

1.2 | Experiments and other tests

Platt (1964) emphasized the importance of experiments that critically distinguish between alternative models and their derived hypotheses when he described the process of strong inference:

- devise alternative hypotheses,
- devise a crucial experiment (or several experiments) each of which will exclude one or more of the hypotheses,
- carry out the experiment(s) carefully to obtain a "clean" result, and
- recycle the procedure with new hypotheses to refine the possibilities (i.e. hypotheses) that remain.

Crucial to Platt's (1964) approach was the idea of multiple competing hypotheses and tests to distinguish between these. What nature should these tests take?

In the dinoflagellate example above, the crucial test of the hypothesis involved a manipulative experiment based on sound principles of experimental design (Chapter 7). Such manipulations provide the strongest inference about our hypotheses and models because we can assess the effects of causal factors on our response variable separately from other factors. James & McCulloch (1985) emphasized that testing biological models, and their subsequent hypotheses, does not occur by simply seeing if their predictions are met in an observational context, although such results offer support for an hypothesis. Along with James & McCulloch (1985), Scheiner (1993), Underwood (1990), Werner (1998), and many others, we argue strongly that manipulative experiments are the best way to properly distinguish between biological models.

There are at least two costs to this strong inference from manipulative experiments. First, experiments nearly always involve some artificial manipulation of nature. The most extreme form of this is when experiments testing some natural process are conducted in the laboratory. Even field experiments will often use artificial structures or mechanisms to implement the manipulation. For example, mesocosms (moderate sized enclosures) are often used to investigate processes happening in large water bodies, although there is evidence from work on lakes that issues related to the small-scale of mesocosms may restrict generalization to whole lakes (Carpenter 1996; see also Resetarits & Fauth 1998). Second, the larger the spatial and temporal scales of the process being investigated, the more difficult it is to meet the guidelines for good experimental design. For example, manipulations of entire ecosystems are crucial for our understanding of the role of natural and anthropogenic disturbances to these systems, especially since natural resource agencies have to manage such systems at this large spatial scale (Carpenter *et al.* 1995). Replication and randomization (two characteristics regarded as important for sensible interpretation of experiments – see Chapter 7) are usually not possible at large scales and novel approaches have been developed to interpret such experiments (Carpenter 1990). The problems of scale and the generality of conclusions from smaller-scale manipulative experiments are challenging issues for experimental biologists (Dunham & Beaupre 1998).

The testing approach on which the methods in this book are based relies on making predictions from our hypothesis and seeing if those predictions apply when observed in a new setting, i.e. with data that were not used to derive the model originally. Ideally, this new setting is experimental at scales relevant for the hypothesis, but this is not always possible. Clearly, there must be additional ways of testing between competing models and their derived hypotheses. Otherwise, disciplines in which experimental manipulation is difficult for practical or ethical reasons, such as meteorology, evolutionary biology, fisheries ecology, etc., could make no scientific progress. The alternative is to predict from our models/hypotheses in new settings that are not

experimentally derived. Hilborn & Mangel (1997), while arguing for experimental studies in ecology where possible, emphasize the approach of "confronting" competing models (or hypotheses) with observational data by assessing how well the data meet the predictions of the model.

Often, the new setting in which we test the predictions of our model may provide us with a contrast of some factor, similar to what we may have set up had we been able to do a manipulative experiment. For example, we may never be able to (nor want to!) test the hypothesis that wildfire in old-growth forests affects populations of forest birds with a manipulative experiment at a realistic spatial scale. However, comparisons of bird populations in forests that have burnt naturally with those that haven't provide a test of the hypothesis. Unfortunately, a test based on such a natural "experiment" (*sensu* Underwood 1990) is weaker inference than a real manipulative experiment because we can never separate the effects of fire from other pre-existing differences between the forests that might also affect bird populations. Assessments of effects of human activities ("environmental impact assessment") are often comparisons of this kind because we can rarely set up a human impact in a truly experimental manner (Downes *et al.* 2001). Well-designed observational (sampling) programs can provide a refutationist test of a null hypothesis (Underwood 1991) by evaluating whether predictions hold, although they cannot demonstrate causality.

While our bias in favor of manipulative experiments is obvious, we hope that we do not appear too dogmatic. Experiments potentially provide the strongest inference about competing hypotheses, but their generality may also be constrained by their artificial nature and limitations of spatial and temporal scale. Testing hypotheses against new observational data provides weaker distinctions between competing hypotheses and the inferential strength of such methods can be improved by combining them with other forms of evidence (anecdotal, mathematical modeling, correlations etc. – see Downes *et al.* 2001, Hilborn & Mangel 1997, McArdle 1996). In practice, most biological investigations will include both observational and experimental approaches. Rigorous and sen-

sible statistical analyses will be relevant at all stages of the investigation.

1.3 | Data, observations and variables

In biology, data usually consist of a collection of observations or objects. These observations are usually sampling units (e.g. quadrats) or experimental units (e.g. individual organisms, aquaria, etc.) and a set of these observations should represent a sample from a clearly defined population (all possible observations in which we are interested). The "actual property measured by the individual observations" (Sokal & Rohlf 1995, p. 9), e.g. length, number of individuals, pH, etc., is called a variable. A random variable (which we will denote as Y, with y being any value of Y) is simply a variable whose values are not known for certain before a sample is taken, i.e. the observed values of a random variable are the results of a random experiment (the sampling process). The set of all possible outcomes of the experiment, e.g. all the possible values of a random variable, is called the sample space. Most variables we deal with in biology are random variables, although predictor variables in models might be fixed in advance and therefore not random. There are two broad categories of random variables: (i) discrete random variables can only take certain, usually integer, values, e.g. the number of cells in a tissue section or number of plants in a forest plot, and (ii) continuous random variables, which take any value, e.g. measurements like length, weight, salinity, blood pressure etc. Kleinbaum *et al.* (1997) distinguish these in terms of "gappiness" – discrete variables have gaps between observations and continuous variables have no gaps between observations.

The distinction between discrete and continuous variables is not always a clear dichotomy; the number of organisms in a sample of mud from a local estuary can take a very large range of values but, of course, must be an integer so is actually a discrete variable. Nonetheless, the distinction between discrete and continuous variables is important, especially when trying to measure uncertainty and probability.

1.4 | Probability

The single most important characteristic of biological data is their uncertainty. For example, if we take two samples, each consisting of the same number of observations, from a population and estimate the mean for some variable, the two means will almost certainly be different, despite the samples coming from the same population. Hilborn & Mangel (1997) proposed two general causes why the two means might be different, i.e. two causes of uncertainty in the expected value of the population. Process uncertainty results from the true population mean being different when the second sample was taken compared with the first. Such temporal changes in biotic variables, even over very short time scales, are common in ecological systems. Observation uncertainty results from sampling error; the mean value in a sample is simply an imperfect estimate of the mean value in the population (all the possible observations) and, because of natural variability between observations, different samples will nearly always produce different means. Observation uncertainty can also result from measurement error, where the measuring device we are using is imperfect. For many biological variables, natural variability is so great that we rarely worry about measurement error, although this might not be the case when the variable is measured using some complex piece of equipment prone to large malfunctions.

In most statistical analyses, we view uncertainty in terms of probabilities and understanding probability is crucial to understanding modern applied statistics. We will only briefly introduce probability here, particularly as it is very important for how we interpret statistical tests of hypotheses. Very readable introductions can be found in Antelman (1997), Barnett (1999), Harrison & Tamaschke (1984) and Hays (1994); from a biological viewpoint in Sokal & Rohlf (1995) and Hilborn & Mangel (1997); and from a philosophical perspective in Mayo (1996).

We usually talk about probabilities in terms of events; the probability of event A occurring is written $P(A)$. Probabilities can be between zero and one; if $P(A)$ equals zero, then the event is

impossible; if $P(A)$ equals one, then the event is certain. As a simple example, and one that is used in nearly every introductory statistics book, imagine the toss of a coin. Most of us would state that the probability of heads is 0.5, but what do we really mean by that statement? The classical interpretation of probability is that it is the relative frequency of an event that we would expect in the long run, or in a long sequence of identical trials. In the coin tossing example, the probability of heads being 0.5 is interpreted as the expected proportion of heads in a long sequence of tosses. Problems with this long-run frequency interpretation of probability include defining what is meant by identical trials and the many situations in which uncertainty has no sensible long-run frequency interpretation, e.g. probability of a horse winning a particular race, probability of it raining tomorrow (Antelman 1997). The long-run frequency interpretation is actually the classical statistical interpretation of probabilities (termed the frequentist approach) and is the interpretation we must place on confidence intervals (Chapter 2) and P values from statistical tests (Chapter 3).

The alternative way of interpreting probabilities is much more subjective and is based on a "degree of belief" about whether an event will occur. It is basically an attempt at quantification of an opinion and includes two slightly different approaches – logical probability developed by Carnap and Jeffreys and subjective probability pioneered by Savage, the latter being a measure of probability specific to the person deriving it. The opinion on which the measure of probability is based may be derived from previous observations, theoretical considerations, knowledge of the particular event under consideration, etc. This approach to probability has been criticized because of its subjective nature but it has been widely applied in the development of prior probabilities in the Bayseian approach to statistical analysis (see below and Chapters 2 and 3).

We will introduce some of the basic rules of probability using a simple biological example with a dichotomous outcome – eutrophication in lakes (e.g. Carpenter *et al.* 1998). Let $P(A)$ be the probability that a lake will go eutrophic. Then $P(\sim A)$ equals one minus $P(A)$, i.e. the probability of not A is one minus the probability of A. In our example, the probability that the lake will not go eutrophic is one minus the probability that it will go eutrophic.

Now consider the $P(B)$, the probability that there will be an increase in nutrient input into the lake. The joint probability of A and B is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (1.1)$$

i.e. the probability that A or B occur [$P(A \cup B)$] is the probability of A plus the probability of B minus the probability of A and B both occurring [$P(A \cap B)$]. In our example, the probability that the lake will go eutrophic or that there will be an increase in nutrient input equals the probability that the lake will go eutrophic plus the probability that the lake will receive increased nutrients minus the probability that the lake will go eutrophic and receive increased nutrients.

These simple rules lead on to conditional probabilities, which are very important in practice. The conditional probability of A, given B, is:

$$P(A \mid B) = P(A \cap B)/P(B) \qquad (1.2)$$

i.e. the probability that A occurs, given that B occurs, equals the probability of A and B both occurring divided by the probability of B occurring. In our example, the probability that the lake will go eutrophic given that it receives increased nutrient input equals the probability that it goes eutrophic and receives increased nutrients divided by the probability that it receives increased nutrients.

We can combine these rules to develop another way of expressing conditional probability – Bayes Theorem (named after the eighteenth-century English mathematician, Thomas Bayes):

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)} \qquad (1.3)$$

This formula allows us to assess the probability of an event A in the light of new information, B. Let's define some terms and then show how this somewhat daunting formula can be useful in practice. $P(A)$ is termed the prior probability of A – it is the probability of A prior to any new information (about B). In our example, it is our probability of a lake going eutrophic, calculated before knowing anything about nutrient inputs, possibly determined from previous studies on eutrophication in

lakes. $P(B|A)$ is the likelihood of B being observed, given that A did occur [a similar interpretation exists for $P(B|\sim A)$]. The likelihood of a model or hypothesis or event is simply the probability of observing some data assuming the model or hypothesis is true or assuming the event occurs. In our example, $P(B|A)$ is the likelihood of seeing a raised level of nutrients, given that the lake has gone eutrophic (A). Finally, $P(A|B)$ is the posterior probability of A, the probability of A after making the observations about B, the probability of a lake going eutrophic after incorporating the information about nutrient input. This is what we are after with a Bayesian analysis, the modification of prior information to posterior information based on a likelihood (Ellison 1996).

Bayes Theorem tells us how probabilities might change based on previous evidence. It also relates two forms of conditional probabilities – the probability of A given B to the probability of B given A. Berry (1996) described this as relating inverse probabilities. Note that, although our simple example used an event (A) that had only two possible outcomes, Bayes formula can also be used for events that have multiple possible outcomes.

In practice, Bayes Theorem is used for estimating parameters of populations and testing hypotheses about those parameters. Equation 1.3 can be simplified considerably (Berry & Stangl 1996, Ellison 1996):

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})} \qquad (1.4)$$

where θ is a parameter to be estimated or an hypothesis to be evaluated, $P(\theta)$ is the "unconditional" prior probability of θ being a particular value, $P(\text{data}|\theta)$ is the likelihood of observing the data if θ is that value, $P(\text{data})$ is the "unconditional" probability of observing the data and is used to ensure the area under the probability distribution of θ equals one (termed "normalization"), and $P(\theta|\text{data})$ is the posterior probability of θ conditional on the data being observed. This formula can be re-expressed in English as:

posterior probability $\propto$ likelihood $\times$
prior probability $\qquad (1.5)$

While we don't advocate a Bayesian philosophy in this book, it is important for biologists to be aware of the approach and to consider it as an alternative way of dealing with conditional probabilities. We will consider the Bayesian approach to estimation in Chapter 2 and to hypothesis testing in Chapter 3.

1.5 | Probability distributions

A random variable will have an associated probability distribution where different values of the variable are on the horizontal axis and the relative probabilities of the possible values of the variable (the sample space) are on the vertical axis. For discrete variables, the probability distribution will comprise a measurable probability for each outcome, e.g. 0.5 for heads and 0.5 for tails in a coin toss, 0.167 for each one of the six sides of a fair die. The sum of these individual probabilities for independent events equals one. Continuous variables are not restricted to integers or any specific values so there are an infinite number of possible outcomes. The probability distribution of a continuous variable (Figure 1.1) is often termed a probability density function (pdf) where the vertical axis is the probability density of the variable [$f(y)$], a rate measuring the probability per unit of the variable at any particular value of the variable (Antelman 1997). We usually talk about the probability associated with a range of values, represented by the area under the probability distribution curve between the two extremes of the range. This area is determined from the integral of the probability density from the lower to the upper value, with the distribution usually normalized so that the total probability under the curve equals one. Note that the probability of any particular value of a continuous random variable is zero because the area under the curve for a single value is zero (Kleinbaum *et al.* 1997) – this is important when we consider the interpretation of probability distributions in statistical hypothesis testing (Chapter 3).

In many of the statistical analyses described in this book, we are dealing with two or more variables and our statistical models will often have more than one parameter. Then we need to switch from single probability distributions to joint

Figure 1.1 Probability distributions for random variables following four common distributions. For the Poisson distribution, we show the distribution for a rare event and a common one, showing the shift of the distribution from skewed to approximately symmetrical.
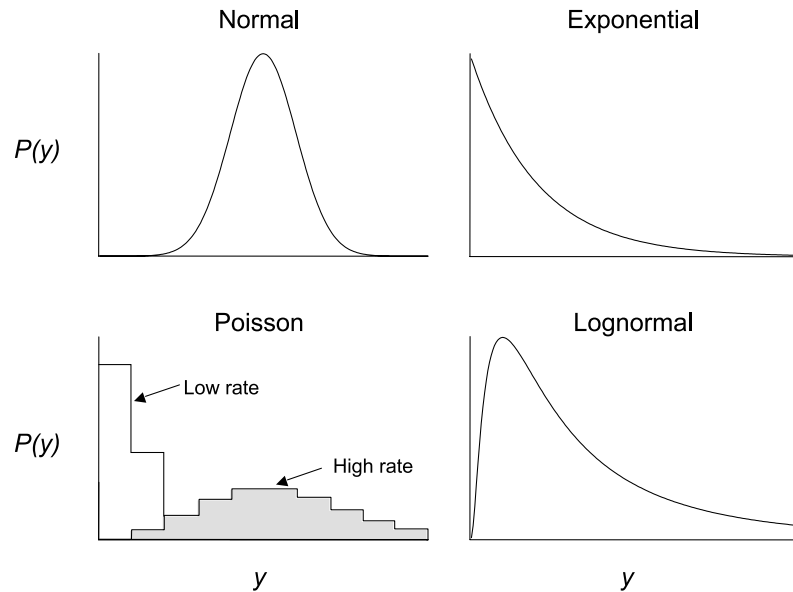

probability distributions where probabilities are measured, not as areas under a single curve, but volumes under a more complex distribution. A common joint pdf is the bivariate normal distribution, to be introduced in Chapter 5.

Probability distributions nearly always refer to the distribution of variables in one or more populations. The expected value of a random variable [$E(Y)$] is simply the mean (μ) of its probability distribution. The expected value is an important concept in applied statistics – most modeling procedures are trying to model the expected value of a random response variable. The mean is a measure of the center of a distribution – other measures include the median (the middle value) and the mode (the most common value). It is also important to be able to measure the spread of a distribution and the most common measures are based on deviations from the center, e.g. the variance is measured as the sum of squared deviations from the mean. We will discuss means and variances, and other measures of the center and spread of distributions, in more detail in Chapter 2.

1.5.1 Distributions for variables

Most statistical procedures rely on knowing the probability distribution of the variable (or the error terms from a statistical model) we are analyzing. There are many probability distributions that we can define mathematically (Evans *et al.* 2000) and some of these adequately describe the distributions of variables in biology. Let's consider continuous variables first.

The normal (also termed Gaussian) distribution is a symmetrical probability distribution with a characteristic bell-shape (Figure 1.1). It is defined as:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(y-\mu)^2/2\sigma^2} \tag{1.6}$$

where $f(y)$ is the probability density of any value y of Y. Note that the normal distribution can be defined simply by the mean (μ) and the variance (σ^2), which are independent of each other. All other terms in the equation are constants. A normal distribution is often abbreviated to $N(Y:\mu,\sigma)$. Since there are infinitely many possible combinations of mean and variance, there is an infinite number of possible normal distributions. The standard normal distribution (z distribution) is a normal distribution with a mean of zero and a variance of one. The normal distribution is the most important probability distribution for data analysis; most commonly used statistical procedures in biology (e.g. linear regression, analysis of variance) assume that the variables being analyzed (or the deviations from a fitted model) follow a normal distribution.

The normal distribution is a symmetrical probability distribution, but continuous variables can have non-symmetrical distributions. Biological variables commonly have a positively skewed distribution, i.e. one with a long right tail (Figure 1.1). One skewed distribution is the lognormal distribution, which means that the logarithm of the