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 relativistic 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 sensible 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:

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

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:

$$P(A|B) = P(A \cap B)/P(B) \quad (1.2)$$

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):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)} \quad (1.3)$$

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):

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

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:

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability} \quad (1.5)$$

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

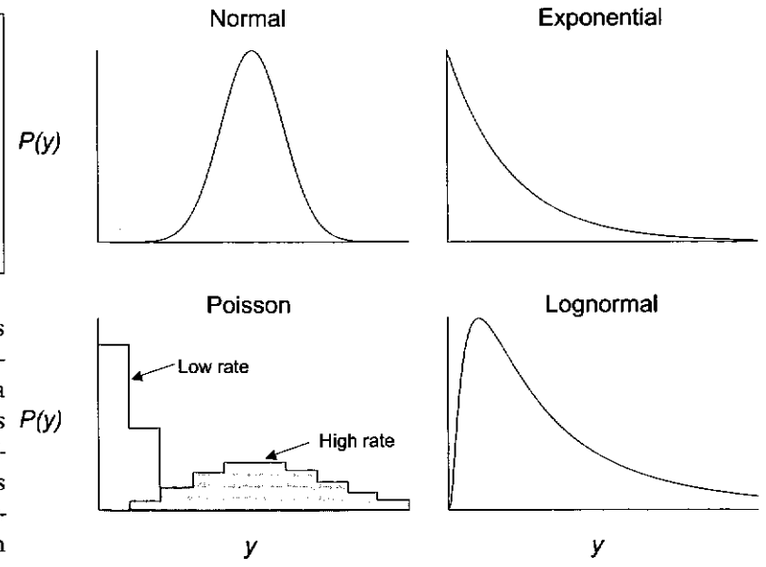

Figure 1.1 Probability distributions for random variables following four common distributions. For the Poisson distribution, we show the distribution for a rare event and a common one, showing the shift of the distribution from skewed to approximately symmetrical.

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:

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

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 variable is normally distributed (suggesting a simple transformation to normality – see Chapter 4). Measurement variables in biology that cannot be less than zero (e.g. length, weight, etc.) often follow lognormal distributions. In skewed distributions like the lognormal, there is a positive relationship between the mean and the variance.

There are some other probability distributions for continuous variables that are occasionally used in specific circumstances. The exponential distribution (Figure 1.1) is another skewed distribution that often applies when the variable is the time to the first occurrence of an event (Fox 1993, Harrison & Tamaschke 1984), such as in failure time analysis. This is a single parameter (λ) distribution with the following probability density function:

$$f(y) = \lambda e^{-\lambda y} \quad (1.7)$$

where $1/\lambda$ is the mean time to first occurrence. Fox (1993) provided some ecological examples.

The exponential and normal distributions are members of the larger family of exponential distributions that can be used as error distributions for a variety of linear models (Chapter 13). Other members of this family include gamma distribution for continuous variables and the binomial and Poisson (see below) for discrete variables.

Two other probability distributions for continuous variables are also encountered (albeit rarely) in biology. The two-parameter Weibull distribution varies between positively skewed and symmetrical depending on parameter values, although versions with three or more parameters are described (Evans *et al.* 2000). This distribution is mainly used for modeling failure rates and times. The beta distribution has two parameters and its shape can range from U to J to symmetrical. The beta distribution is commonly used as a prior probability distribution for dichotomous variables in Bayesian analyses (Evans *et al.* 2000).

There are also probability distributions for discrete variables. If we toss a coin, there are two possible outcomes – heads or tails. Processes with only two possible outcomes are common in biology, e.g. animals in an experiment can either live or die, a particular species of tree can be either present or absent from samples from a forest. A process that can only have one of two outcomes is sometimes called a Bernoulli trial and we often call the two possible outcomes success and failure. We will only consider a stationary Bernoulli trial, which is one where the probability of success is the same for each trial, i.e. the trials are independent.

The probability distribution of the number of successes in n independent Bernoulli trials is called the binomial distribution, a very important probability distribution in biology:

$$P(y=r) = \frac{n!}{r!(n-r)!} \pi^r (1-\pi)^{n-r} \quad (1.8)$$

where $P(y=r)$ is the probability of a particular value (y) of the random variable (Y) being r successes out of n trials, n is the number of trials and π is the probability of a success. Note that n, the number of trials is fixed, and therefore the value of a binomial random variable cannot exceed n. The binomial distribution can be used to calculate probabilities for different numbers of successes out of n trials, given a known probability of success on any individual trial. It is also important as an error distribution for modeling variables with binary outcomes using logistic regression (Chapter 13). A generalization of the binomial distribution to when there are more than two possible outcomes is the multinomial distribution, which is the joint probability distribution of multiple outcomes from n fixed trials.

Another very important probability distribution for discrete variables is the Poisson distribution, which usually describes variables representing the number of (usually rare) occurrences of a particular event in an interval of time or space, i.e. counts. For example, the number of organisms in a plot, the number of cells in a microscope field of view, the number of seeds taken by a bird per minute. The probability distribution of a Poisson variable is:

$$P(y=r) = \frac{e^{-\mu}\mu^r}{r!} \quad (1.9)$$

where $P(y=r)$ is the probability that the number of occurrences of an event (y) equals an integer value ($r=0, 1, 2\ldots$), μ is the mean (and variance) of the number of occurrences. A Poisson variable can take any integer value between zero and infinity because the number of trials, in contrast to the

binomial and the multinomial, is not fixed. One of the characteristics of a Poisson distribution is that the mean (μ) equals the variance (σ^2). For small values of μ, the Poisson distribution is positively skewed but once μ is greater than about five, the distribution is symmetrical (Figure 1.1).

The Poisson distribution has a wide range of applications in biology. It actually describes the occurrence of random events in space (or time) and has been used to examine whether organisms have random distributions in nature (Ludwig & Reynolds 1988). It also has wide application in many applied statistical procedures, e.g. counts in cells in contingency tables are often assumed to be Poisson random variables and therefore a Poisson probability distribution is used for the error terms in log-linear modeling of contingency tables (Chapter 14).

A simple example might help in understanding the difference between the binomial and the Poisson distributions. If we know the average number of seedlings of mountain ash trees (*Eucalyptus regnans*) per plot in some habitat, we can use the Poisson distribution to model the probability of different numbers of seedlings per plot, assuming independent sampling. The binomial distribution would be used if we wished to model the number of plots with seedlings out of a fixed number of plots, knowing the probability of a plot having a seedling.

Another useful probability distribution for counts is the negative binomial (White & Bennetts 1996). It is defined by two parameters, the mean and a dispersion parameter, which measures the degree of "clumping" in the distribution. White & Bennetts (1996) pointed out that the negative binomial has two potential advantages over the Poisson for representing skewed distributions of counts of organisms: (i) the mean does not have to equal the variance, and (ii) independence of trials (samples) is not required (see also Chapter 13).

These probability distributions are very important in data analysis. We can test whether a particular variable follows one of these distributions by calculating the expected frequencies and comparing them to observed frequencies with a goodness-of-fit test (Chapter 14). More importantly, we can model the expected value of a response variable [$E(Y)$] against a range of predictor (independent) variables if we know the probability distribution of our response variable.

1.5.2 Distributions for statistics

The remaining theoretical distributions to examine are those used for determining probabilities of sample statistics, or modifications thereof. These distributions are used extensively for estimation and hypothesis testing. Four particularly important ones are as follows.

1. The z or normal distribution represents the probability distribution of a random variable that is the ratio of the difference between a sample statistic and its population value to the standard deviation of the population statistic (Figure 1.2).

2. Student's t distribution (Figure 1.2) represents the probability distribution of a random variable that is the ratio of the difference between a sample statistic and its population value to the standard deviation of the distribution of the sample statistic. The t distribution is a symmetrical distribution very similar to a normal distribution, bounded by infinity in both directions. Its shape becomes more similar with increasing sample size (Figure 1.2). We can convert a single sample statistic to a t value and use the t distribution to determine the probability of obtaining that t value (or one smaller or larger) for a specified value of the population parameter (Chapters 2 and 3).

3. χ^2 (chi-square) distribution (Figure 1.2) represents the probability distribution of a variable that is the square of values from a standard normal distribution (Section 1.5). Values from a χ^2 distribution are bounded by zero and infinity. Variances have a χ^2 distribution so this distribution is used for interval estimation of population variances (Chapter 2). We can also use the χ^2 distribution to determine the probability of obtaining a sample difference (or one smaller or larger) between observed values and those predicted by a model (Chapters 13 and 14).

4. F distribution (Figure 1.2) represents the probability distribution of a variable that is the ratio of two independent χ^2 variables, each

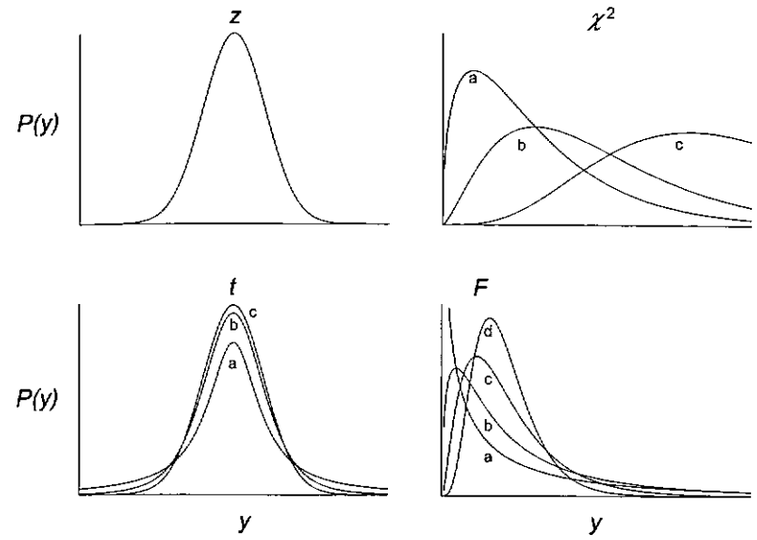

Figure 1.2 Probability distributions for four common statistics. For the t, χ^2, and F distributions, we show distributions for three or four different degrees of freedom (a to d, in increasing order), to show how the shapes of these distributions change.

divided by its df (degrees of freedom) (Hays 1994). Because variances are distributed as χ^2, the F distribution is used for testing hypotheses about ratios of variances. Values from the F distribution are bounded by zero and infinity. We can use the F distribution to determine the probability of obtaining a sample variance ratio (or one larger) for a specified value of the true ratio between variances (Chapters 5 onwards).

All four distributions have mathematical derivations that are too complex to be of much interest to biologists (see Evans *et al.* 2000). However, these distributions are tabled in many textbooks and programmed into most statistical software, so probabilities of obtaining values from each, within a specific range, can be determined. These distributions are used to represent the probability distributions of the sample statistics (z, t, χ^2 or F) that we would expect from repeated random sampling from a population or populations. Different versions of each distribution are used depending on the degrees of freedom associated with the sample or samples (see Box 2.1 and Figure 1.2).

Chapter 2

Estimation

2.1 Samples and populations

Biologists usually wish to make inferences (draw conclusions) about a population, which is defined as the collection of all the *possible* observations of interest. Note that this is a statistical population, not a biological population (see below). The collection of observations we take from the population is called a sample and the number of observations in the sample is called the sample size (usually given the symbol n). Measured characteristics of the sample are called statistics (e.g. sample mean) and characteristics of the population are called parameters (e.g. population mean).

The basic method of collecting the observations in a sample is called simple random sampling. This is where any observation has the same probability of being collected, e.g. giving every rat in a holding pen a number and choosing a sample of rats to use in an experiment with a random number table. We rarely sample truly randomly in biology, often relying on haphazard sampling for practical reasons. The aim is always to sample in a manner that doesn't create a bias in favour of any observation being selected. Other types of sampling that take into account heterogeneity in the population (e.g. stratified sampling) are described in Chapter 7. Nearly all applied statistical procedures that are concerned with using samples to make inferences (i.e. draw conclusions) about populations assume some form of random sampling. If the sampling is not random, then we are never sure quite what population is represented by our sample. When random sampling from clearly defined populations is not possible, then interpretation of standard methods of estimation becomes more difficult.

Populations must be defined at the start of any study and this definition should include the spatial and temporal limits to the population and hence the spatial and temporal limits to our inference. Our formal statistical inference is restricted to these limits. For example, if we sample from a population of animals at a certain location in December 1996, then our inference is restricted to that location in December 1996. We cannot infer what the population might be like at any other time or in any other place, although we can speculate or make predictions.

One of the reasons why classical statistics has such an important role in the biological sciences, particularly agriculture, botany, ecology, zoology, etc., is that we can often define a population about which we wish to make inferences and from which we can sample randomly (or at least haphazardly). Sometimes the statistical population is also a biological population (a group of individuals of the same species). The reality of random sampling makes biology a little different from other disciplines that use statistical analyses for inference. For example, it is often difficult for psychologists or epidemiologists to sample randomly because they have to deal with whatever subjects or patients are available (or volunteer!).

The main reason for sampling randomly from a clearly defined population is to use sample statistics (e.g. sample mean or variance) to estimate population parameters of interest (e.g. population mean or variance). The population parameters cannot be measured directly because the populations are usually too large, i.e. they contain too many observations for practical measurement. It is important to remember that population parameters are usually considered to be fixed, but unknown, values so they are not random variables and do not have probability distributions. Note that this contrasts with the Bayesian approach where population parameters are viewed as random variables (Section 2.6). Sample statistics are random variables, because their values depend on the outcome of the sampling experiment, and therefore they do have probability distributions, called sampling distributions.

What are we after when we estimate population parameters? A good estimator of a population parameter should have the following characteristics (Harrison & Tamaschke 1984, Hays 1994).

- It should be unbiased, meaning that the expected value of the sample statistic (the mean of its probability distribution) should equal the parameter. Repeated samples should produce estimates which do not consistently under- or over-estimate the population parameter.
- It should be consistent so as the sample size increases then the estimator will get closer to the population parameter. Once the sample includes the whole population, the sample statistic will obviously equal the population parameter, by definition.
- It should be efficient, meaning it has the lowest variance among all competing estimators. For example, the sample mean is a more efficient estimator of the population mean of a variable with a normal probability distribution than the sample median, despite the two statistics being numerically equivalent.

There are two broad types of estimation:

1. point estimates provide a single value which estimates a population parameter, and
2. interval estimates provide a range of values that might include the parameter with a known probability, e.g. confidence intervals.

Later in this chapter we discuss different methods of estimating parameters, but, for now, let's consider some common population parameters and their point estimates.

2.2 Common parameters and statistics

Consider a population of observations of the variable Y measured on all N sampling units in the population. We take a random sample of n observations ($y_1, y_2, y_3, \ldots y_i, \ldots y_n$) from the population. We usually would like information about two aspects of the population, some measure of location or central tendency (i.e. where is the middle of the population?) and some measure of the spread (i.e. how different are the observations in the population?). Common estimates of parameters of location and spread are given in Table 2.1 and illustrated in Box 2.2.

2.2.1 Center (location) of distribution

Estimators for the center of a distribution can be classified into three general classes, or broad types (Huber 1981, Jackson 1986). First are L-estimators, based on the sample data being ordered from smallest to largest (order statistics) and then forming a linear combination of weighted order statistics. The sample mean ($\bar{y}$), which is an unbiased estimator of the population mean (μ), is an L-estimator where each observation is weighted by $1/n$ (Table 2.1). Other common L-estimators include the following.

- The median is the middle measurement of a set of data. Arrange the data in order of magnitude (i.e. ranks) and weight all observations except the middle one by zero. The median is an unbiased estimator of the population mean for normal distributions, is a better estimator of the center of skewed distributions and is more resistant to outliers (extreme values very different to the rest of the sample; see Chapter 4).
- The trimmed mean is the mean calculated after omitting a proportion (commonly 5%) of the highest (and lowest) observations, usually to deal with outliers.
- The Winsorized mean is determined as for trimmed means except the omitted observations are replaced by the nearest remaining value.

Second are M-estimators, where the weightings given to the different observations change

Table 2.1 Common population parameters and sample statistics

Parameter	Statistic	Formula
Mean (μ)	$\bar{y}$	$\frac{\sum_{i=1}^{n} y_i}{n}$
Median	Sample median	$y_{(n+1)/2}$ if n odd $(y_{n/2} + y_{(n/2)+1})/2$ if n even
Variance (σ^2)	s^2	$\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$
Standard deviation (σ)	s	$\sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}}$
Median absolute deviation (MAD)	Sample MAD	median$[\lvert y_i - \text{median} \rvert]$
Coefficient of variation (CV)	Sample CV	$\frac{s}{\bar{y}} \times 100$
Standard error of $\bar{y}$ ($\sigma_{\bar{y}}$)	$s_{\bar{y}}$	$\frac{s}{\sqrt{n}}$
95% confidence interval for μ		$\bar{y} - t_{0.05(n-1)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{y} + t_{0.05(n-1)}\frac{s}{\sqrt{n}}$

gradually from the middle of the sample and incorporate a measure of variability in the estimation procedure. They include the Huber M-estimator and the Hampel M-estimator, which use different functions to weight the observations. They are tedious to calculate, requiring iterative procedures, but maybe useful when outliers are present because they downweight extreme values. They are not commonly used but do have a role in robust regression and ANOVA techniques for analyzing linear models (regression in Chapter 5 and ANOVA in Chapter 8).

Finally, R-estimators are based on the ranks of the observations rather than the observations themselves and form the basis for many rank-based "non-parametric" tests (Chapter 3). The only common R-estimator is the Hodges–Lehmann estimator, which is the median of the averages of all possible pairs of observations.

For data with outliers, the median and trimmed or Winsorized means are the simplest to calculate although these and M- and R-estimators are now commonly available in statistical software.

2.2.2 Spread or variability

Various measures of the spread in a sample are provided in Table 2.1. The range, which is the difference between the largest and smallest observation, is the simplest measure of spread, but there is no clear link between the sample range and the population range and, in general, the range will rise as sample size increases. The sample variance, which estimates the population variance, is an important measure of variability in many statistical analyses. The numerator of the formula is called the sum of squares (SS, the sum of squared deviations of each observation from the sample mean) and the variance is the average of these squared deviations. Note that we might expect to divide by n to calculate an average, but then s^2 consistently underestimates σ^2 (i.e. it is biased), so we divide by $n-1$ to make s^2 an unbiased estimator of σ^2. The one difficulty with s^2 is that its units are the square of the original observations, e.g. if the observations are lengths in mm, then the variance is in mm^2, an area not a length.

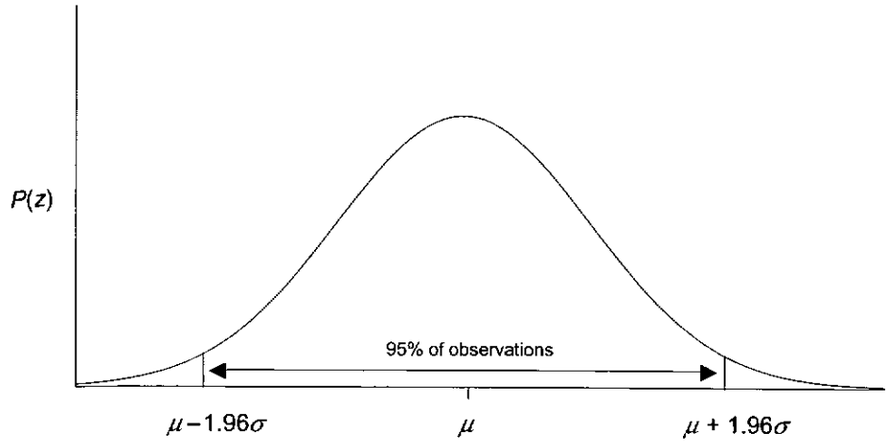

Figure 2.1 Plot of normal probability distribution, showing points between which values 95% of all values occur.

The sample standard deviation, which estimates σ, the population standard deviation, is the square root of the variance. In contrast to the variance, the standard deviation is in the same units as the original observations.

The coefficient of variation (CV) is used to compare standard deviations between populations with different means and it provides a measure of variation that is independent of the measurement units. The sample coefficient of variation CV describes the standard deviation as a percentage of the mean; it estimates the population CV.

Some measures of spread that are more robust to unusual observations include the following.

- The median absolute deviation (MAD) is less sensitive to outliers than the above measures and is the sensible measure of spread to present in association with medians.
- The interquartile range is the difference between the first quartile (the observation which has 0.25 or 25% of the observations below it) and the third quartile (the observation which has 0.25 of the observations above it). It is used in the construction of boxplots (Chapter 4).

For some of these statistics (especially the variance and standard deviation), there are equivalent formulae that can be found in any statistics textbook that are easier to use with a hand calculator. We assume that, in practice, biologists will use statistical software to calculate these statistics and, since the alternative formulae do not assist in the understanding of the concepts, we do not provide them.

2.3 Standard errors and confidence intervals for the mean

2.3.1 Normal distributions and the Central Limit Theorem

Having an estimate of a parameter is only the first step in estimation. We also need to know how precise our estimate is. Our estimator may be the most precise of all the possible estimators, but if its value still varies widely under repeated sampling, it will not be very useful for inference. If repeated sampling produces an estimator that is very consistent, then it is precise and we can be confident that it is close to the parameter (assuming that it is unbiased). The traditional logic for determining precision of estimators is well covered in almost every introductory statistics and biostatistics book (we strongly recommend Sokal & Rohlf 1995), so we will describe it only briefly, using normally distributed variables as an example.

Assume that our sample has come from a normally distributed population (Figure 2.1). For any normal distribution, we can easily determine what proportions of observations in the

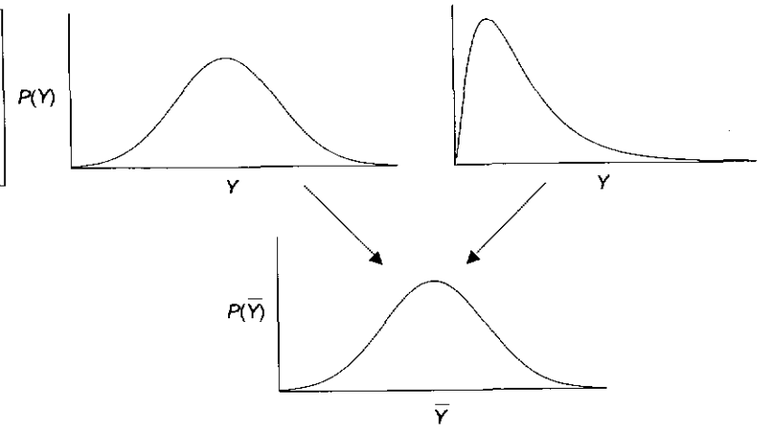

Figure 2.2 Illustration of the principle of the Central Limit Theorem, where repeated samples with large n from any distribution will have sample means with a normal distribution.

population occur within certain distances from the mean:

- 50% of population falls between $\mu \pm 0.674\sigma$
- 95% of population falls between $\mu \pm 1.960\sigma$
- 99% of population falls between $\mu \pm 2.576\sigma$.

Therefore, if we know μ and σ, we can work out these proportions for any normal distribution. These proportions have been calculated and tabulated in most textbooks, but only for the standard normal distribution, which has a mean of zero and a standard deviation (or variance) of one. To use these tables, we must be able to transform our sample observations to their equivalent values in the standard normal distribution. To do this, we calculate deviations from the mean in standard deviation units:

$$z = \frac{y_i - \mu}{\sigma} \tag{2.1}$$

These deviations are called normal deviates or standard scores. This z transformation in effect converts any normal distribution to the standard normal distribution.

Usually we only deal with a single sample (with n observations) from a population. If we took many samples from a population and calculated all their sample means, we could plot the frequency (probability) distribution of the sample means (remember that the sample mean is a random variable). This probability distribution is called the sampling distribution of the mean and has three important characteristics.

- The probability distribution of means of samples from a normal distribution is also normally distributed.
- As the sample size increases, the probability distribution of means of samples from any distribution will approach a normal distribution. This result is the basis of the Central Limit Theorem (Figure 2.2).
- The expected value or mean of the probability distribution of sample means equals the mean of the population (μ) from which the samples were taken.

2.3.2 Standard error of the sample mean

If we consider the sample means to have a normal probability distribution, we can calculate the variance and standard deviation of the sample means, just like we could calculate the variance of the observations in a single sample. The expected value of the standard deviation of the sample means is:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \tag{2.2}$$

where σ is the standard deviation of the original population from which the repeated samples were taken and n is the size of samples.

We are rarely in the position of having many samples from the same population, so we estimate the standard deviation of the sample means from our single sample. The standard deviation of the sample means is called the standard error of the mean:

$$s_{\bar{y}} = \frac{s}{\sqrt{n}} \tag{2.3}$$

where s is the sample estimate of the standard deviation of the original population and n is the sample size.

The standard error of the mean is telling us about the variation in our sample mean. It is termed "error" because it is telling us about the error in using $\bar{y}$ to estimate μ (Snedecor & Cochran 1989). If the standard error is large, repeated samples would likely produce very different means, and the mean of any single sample might not be close to the true population mean. We would not have much confidence that any specific sample mean is a good estimate of the population mean. If the standard error is small, repeated samples would likely produce similar means, and the mean of any single sample is more likely to be close to the true population mean. Therefore, we would be quite confident that any specific sample mean is a good estimate of the population mean.

2.3.3 Confidence intervals for population mean

In Equation 2.1, we converted any value from a normal distribution into its equivalent value from a standard normal distribution, the z score. Equivalently, we can convert any sample mean into its equivalent value from a standard normal distribution of means using:

$$z = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}} \tag{2.4}$$

where the denominator is simply the standard deviation of the mean, $\sigma/\sqrt{n}$, or standard error. Because this z score has a normal distribution, we can determine how confident we are in the sample mean, i.e. how close it is to the true population mean (the mean of the distribution of sample means). We simply determine values in our distribution of sample means between which a given percentage (often 95% by convention) of means occurs, i.e. between which values of $(\bar{y} - \mu)/\sigma_{\bar{y}}$ do 95% of values lie? As we showed above, 95% of a normal distribution falls between $\mu \pm 1.960\sigma$, so 95% of sample means fall between $\mu \pm 1.96\sigma_{\bar{y}}$ (1.96 times the standard deviation of the distribution of sample means, the standard error).

Now we can combine this information to make a confidence interval for μ:

$$P\{\bar{y} - 1.96\sigma_{\bar{y}} \leq \mu \leq \bar{y} + 1.96\sigma_{\bar{y}}\} = 0.95 \tag{2.5}$$

This confidence interval is an interval estimate for the population mean, although the probability statement is actually about the interval, not about the population parameter, which is fixed. We will discuss the interpretation of confidence intervals in the next section. The only problem is that we very rarely know σ in practice, so we never actually know $\sigma_{\bar{y}}$; we can only estimate the standard error from s (sample standard deviation). Our standard normal distribution of sample means is now the distribution of $(\bar{y} - \mu)/s_{\bar{y}}$. This is a random variable called t and it has a probability distribution that is not quite normal. It follows a t distribution (Chapter 1), which is flatter and more spread than a normal distribution. Therefore, we must use the t distribution to calculate confidence intervals for the population mean in the common situation of not knowing the population standard deviation.

The t distribution (Figure 1.2) is a symmetrical probability distribution centered around zero and, like a normal distribution, it can be defined mathematically. Proportions (probabilities) for a standard t distribution (with a mean of zero and standard deviation of one) are tabled in most statistics books. In contrast to a normal distribution, however, t has a slightly different distribution depending on the sample size (well, for mathematical reasons, we define the different t distributions by $n-1$, called the degrees of freedom (df) (see Box 2.1), rather than n). This is because s provides an imprecise estimate of σ if the sample size is small, increasing in precision as the sample size increases. When n is large (say >30), the t distribution is very similar to a normal distribution (because our estimate of the standard error based on s will be very close to the real standard error). Remember, the z distribution is simply the probability distribution of $(y - \mu)/\sigma$ or $(\bar{y} - \mu)/\sigma_{\bar{y}}$ if we are dealing with sample means. The t distribution is simply the probability distribution of $(\bar{y} - \mu)/s_{\bar{y}}$ and there is a different t distribution for each df $(n-1)$.

The confidence interval (95% or 0.95) for the population mean then is:

$$P\{\bar{y} - t_{0.05(n-1)}s_{\bar{y}} \leq \mu \leq \bar{y} + t_{0.05(n-1)}s_{\bar{y}}\} = 0.95 \tag{2.6}$$

where $t_{0.05(n-1)}$ is the value from the t distribution with $n-1$ df between which 95% of all t values lie and $s_{\bar{y}}$ is the standard error of the mean. Note that the size of the interval will depend on the sample size and the standard deviation of the sample, both of which are used to calculate the standard

Box 2.1 | Explanation of degrees of freedom

Degrees of freedom (df) is one of those terms that biologists use all the time in statistical analyses but few probably really understand. We will attempt to make it a little clearer. The degrees of freedom is simply the number of observations in our sample that are "free to vary" when we are estimating the variance (Harrison & Tamaschke 1984). Since we have already determined the mean, then only $n-1$ observations are free to vary because knowing the mean and $n-1$ observations, the last observation is fixed. A simple example – say we have a sample of observations, with values 3, 4 and 5. We know the sample mean (4) and we wish to estimate the variance. Knowing the mean and one of the observations doesn't tell us what the other two must be. But if we know the mean and two of the observations (e.g. 3 and 4), the final observation is fixed (it must be 5). So, knowing the mean, only two observations ($n-1$) are free to vary. As a general rule, the df is the number of observations minus the number of parameters included in the formula for the variance (Harrison & Tamaschke 1984).

error, and also on the level of confidence we require (Box 2.3).

We can use Equation 2.6 to determine confidence intervals for different levels of confidence, e.g. for 99% confidence intervals, simply use the t value between which 99% of all t values lie. The 99% confidence interval will be wider than the 95% confidence interval (Box 2.3).

2.3.4 Interpretation of confidence intervals for population mean

It is very important to remember that we usually do not consider μ a random variable but a fixed, albeit unknown, parameter and therefore the confidence interval is not a probability statement about the population mean. We are not saying there is a 95% probability that μ falls within this specific interval that we have determined from our sample data; μ is fixed, so this confidence interval we have calculated for a single sample either contains μ or it doesn't. The probability associated with confidence intervals is interpreted as a long-run frequency, as discussed in Chapter 1. Different random samples from the same population will give different confidence intervals and if we took 100 samples of this size (n), and calculated the 95% confidence interval from each sample, 95 of the intervals would contain μ and five wouldn't. Antelman (1997, p. 375) summarizes a confidence interval succinctly as ". . . one interval generated by a procedure that will give correct intervals 95% of the time".

2.3.5 Standard errors for other statistics

The standard error is simply the standard deviation of the probability distribution of a specific statistic, such as the mean. We can, however, calculate standard errors for other statistics besides the mean. Sokal & Rohlf (1995) have listed the formulae for standard errors for many different statistics but noted that they might only apply for large sample sizes or when the population from which the sample came was normal. We can use the methods just described to reliably determine standard errors for statistics (and confidence intervals for the associated parameters) from a range of analyses that assume normality, e.g. regression coefficients. These statistics, when divided by their standard error, follow a t distribution and, as such, confidence intervals can be determined for these statistics (confidence interval = $t \times$ standard error).

When we are not sure about the distribution of a sample statistic, or know that its distribution is non-normal, then it is probably better to use resampling methods to generate standard errors (Section 2.5). One important exception is the sample variance, which has a known distribution that is not normal, i.e. the Central Limit Theorem does not apply to variances. To calculate confidence intervals for the population variance, we need to use the chi-square (χ^2) distribution, which is the distribution of the following random variable:

$$\chi^2 = \frac{(y-\mu)^2}{\sigma^2} \qquad (2.7)$$

Box 2.2 | Worked example of estimation: chemistry of forested watersheds

Lovett *et al.* (2000) studied the chemistry of forested watersheds in the Catskill Mountains in New York State. They chose 39 sites (observations) on first and second order streams and measured the concentrations of ten chemical variables (NO_3^-, total organic N, total N, NH_4^-, dissolved organic C, SO_4^{2-}, Cl^-, Ca^{2+}, Mg^{2+}, H^+), averaged over three years, and four watershed variables (maximum elevation, sample elevation, length of stream, watershed area). We will assume that the 39 sites represent a random sample of possible sites in the central Catskills and will focus on point estimation for location and spread of the populations for two variables, SO_4^{2-} and Cl^-, and interval estimation for the population mean of these two variables. We also created a modified version of SO_4^{2-} where we replaced the largest value (72.1 μmol l^{-1} at site BWS6) by an extreme value of 200 μmol l^{-1} to illustrate the robustness of various statistics to outliers.

Boxplots (Chapter 4) for both variables are presented in Figure 4.3. Note that SO_4^{2-} has a symmetrical distribution whereas Cl^- is positively skewed with outliers (values very different from rest of sample). Summary statistics for SO_4^{2-} (original and modified) and Cl^- are presented below.

Estimate	SO_4^{2-}	Modified SO_4^{2-}	Cl^-
Mean	61.92	65.20	22.84
Median	62.10	62.10	20.50
5% trimmed mean	61.90	61.90	20.68
Huber's M-estimate	61.67	61.67	20.21
Hampel's M-estimate	61.85	61.62	19.92
Standard deviation	5.24	22.70	12.38
Interquartile range	8.30	8.30	7.80
Median absolute deviation	4.30	4.30	3.90
Standard error of mean	0.84	3.64	1.98
95% confidence interval for mean	60.22–63.62	57.84–72.56	18.83–26.86

Given the symmetrical distribution of SO_4^{2-}, the mean and median are similar as expected. In contrast, the mean and the median are different by more than two units for Cl^-, as we would expect for a skewed distribution. The median is a more reliable estimator of the center of the skewed distribution for Cl^-, and the various robust estimates of location (median, 5% trimmed mean, Huber's and Hampel's M-estimates) all give similar values. The standard deviation for Cl^- is also affected by the outliers, and the confidence intervals are relatively wide.

The modified version of SO_4^{2-} also shows the sensitivity of the mean and the standard deviation to outliers. Of the robust estimators for location, only Hampel's M-estimate changes marginally, whereas the mean changes by more than three units. Similarly, the standard deviation (and therefore the standard error and 95%

confidence interval) is much greater for the modified variable, whereas the interquartile range and the median absolute deviation are unaffected by the outlier.

We also calculated bootstrap estimates for the mean and the median of SO_4^{2-} concentrations, based on 1000 bootstrap samples ($n = 39$) with replacement from the original sample of 39 sites. The bootstrap estimate was the mean of the 1000 bootstrap sample statistics, the bootstrap standard error was the standard deviation of the 1000 bootstrap sample statistics and the 95% confidence interval was determined from 25th and 975th values of the bootstrap statistics arranged in ascending order. The two estimates of the mean were almost identical, and although the standard error was smaller for the usual method, the percentile 95% confidence interval for the bootstrap method was narrower. The two estimates for the median were identical, but the bootstrap method allows us to estimate a standard error and a confidence interval.

	Usual	Bootstrap
Mean	61.92	61.91
Standard error	0.84	0.88
95% confidence interval	60.22–63.62	60.36–63.59
Median	61.72	61.72
Standard error	NA	1.34
95% confidence interval	NA	58.60–63.40

The frequency distributions of the bootstrap means and medians are presented in Figure 2.4. The distribution of bootstrap means is symmetrical whereas the bootstrap distribution of medians is skewed. This is commonly the case and the confidence interval for the median is not symmetrical around the bootstrap estimate. We also calculated the bias corrected bootstrap confidence intervals. Forty nine percent of bootstrap means were below the bootstrap estimate of 61.91, so the bias-corrected confidence interval is basically the same as the standard bootstrap. Forty four percent of bootstrap medians were below the bootstrap estimate of 61.72, so $z_0 = -0.151$ and $(2z_0 + 1.96) = 1.658$ and $(2z_0 - 1.96) = -2.262$. The percentiles, from the normal cumulative distribution, are 95.2% (upper) and 1.2% (lower). However, because so many of the bootstrap medians were the same value, these bias-corrected percentiles did not change the confidence intervals.

This is simply the square of the standard z score discussed above (see also Chapter 1). Because we square the numerator, χ^2 is always positive, ranging from zero to ∞. The χ^2 distribution is a sampling distribution so, like the random variable t, there are different probability distributions for χ^2 for different sample sizes; this is reflected in the degrees of freedom ($n-1$). For small df, the probability distribution is skewed to the right (Figure 1.2) but it approaches normality as df increases.

Now back to the sample variance. It turns out that the probability distribution of the sample variance is a chi-square distribution. Strictly speaking,

$$\frac{(n-1)s^2}{\sigma^2} \qquad (2.8)$$

is distributed as χ^2 with $n-1$ df (Hays 1994). We can rearrange Equation 2.8, using the chi-square distribution, to determine a confidence interval for the variance:

$$P\left\{\frac{s^2(n-1)}{\chi^2_{n-1}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{n-1}}\right\} = 0.95 \qquad (2.9)$$

where the lower bound uses the χ^2 value below which 2.5% of all χ^2 values fall and the upper bound uses the χ^2 value above which 2.5% of all χ^2 values fall. Remember the long-run frequency interpretation of this confidence interval – repeated sampling would result in confidence intervals of which 95% would include the true population variance. Confidence intervals on variances are very important for the interpretation of variance components in linear models (Chapter 8).

Box 2.3 Effect of different sample variances, sample sizes and degrees of confidence on confidence interval for the population mean

We will again use the data from Lovett *et al.* (2000) on the chemistry of forested watersheds in the Catskill Mountains in New York State and focus on interval estimation for the mean concentration of SO_4^{2-} in all the possible sites that could have been sampled.

Original sample

Sample ($n = 39$) with a mean concentration of SO_4^{2-} of 61.92 and s of 5.24. The t value for 95% confidence intervals with 38 df is 2.02. The 95% confidence interval for population mean SO_4^{2-} is 60.22 – 63.62, i.e. 3.40.

Different sample variance

Sample ($n = 39$) with a mean concentration of SO_4^{2-} of 61.92 and s of 10.48 (twice original). The t value for 95% confidence intervals with 38 df is 2.02. The 95% confidence interval for population mean SO_4^{2-} is 58.53 – 65.31, i.e. 6.78 (cf. 3.40).

So more variability in population (and sample) results in a wider confidence interval.

Different sample size

Sample ($n = 20$; half original) with a mean concentration of SO_4^{2-} of 61.92 and s of 5.24. The t value for 95% confidence intervals with 19 df is 2.09. The 95% confidence interval for population mean SO_4^{2-} is 59.47 – 64.37, i.e. 4.90 (cf. 3.40).

So a smaller sample size results in wider interval because our estimates of s and $s_{\bar{y}}$ are less precise.

Different level of confidence (99%)

Sample ($n = 39$) with a mean concentration of SO_4^{2-} of 61.92 and s of 5.24. The t value for 99% confidence intervals with 38 df is 2.71. The 95% confidence interval for population mean SO_4^{2-} is 59.65 – 64.20, i.e. 4.55 (cf. 3.40).

So requiring a greater level of confidence results in a wider interval for a given n and s.

2.4 Methods for estimating parameters

2.4.1 Maximum likelihood (ML)

A general method for calculating statistics that estimate specific parameters is called Maximum Likelihood (ML). The estimates of population parameters (e.g. the population mean) provided earlier in this chapter are ML estimates, except for the variance where we correct the estimate to reduce bias. The logic of ML estimation is deceptively simple. Given a sample of observations from a population, we find estimates of one (or more) parameter(s) that maximise the likelihood of observing those data. To determine maximum likelihood estimators, we need to appreciate the likelihood function, which provides the likelihood of the observed data (and therefore our sample statistic) for all possible values of the parameter we are trying to estimate. For example, imagine we have a sample of observations with a sample mean of $\bar{y}$. The likelihood function, assuming a normal distribution and for a given standard

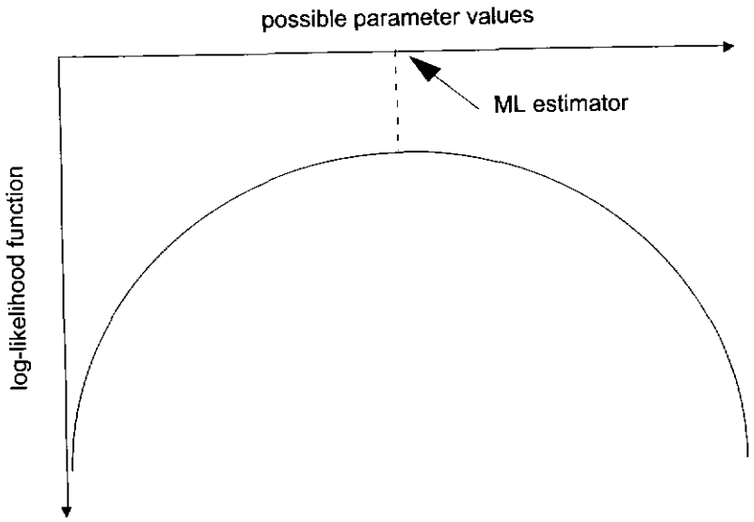

Figure 2.3 Generalized log-likelihood function for estimating a parameter.

deviation, is the likelihood of observing the data for all possible values of μ, the population mean. In general, for a parameter θ, the likelihood function is:

$$L(y;\theta) = \prod_{i=1}^{n} f(y_i;\theta) \qquad (2.10)$$

where $f(y_i;\theta)$ is the joint probability distribution of y_i and θ, i.e. the probability distribution of Y for possible values of θ. In many common situations, $f(y_i;\theta)$ is a normal probability distribution. The ML estimator of θ is the one that maximizes this likelihood function. Working with products (Π) in Equation 2.10 is actually difficult in terms of computation so it is more common to maximize the log-likelihood function:

$$L(\theta) = \ln\left[\prod_{i=1}^{n} f(y_i;\theta)\right] = \sum_{i=1}^{n} \ln[f(y_i;\theta)] \qquad (2.11)$$

For example, the ML estimator of μ (knowing σ^2) for a given sample is the value of μ which maximises the likelihood of observing the data in the sample. If we are trying to estimate μ from a normal distribution, then the $f(y_i;\mu)$ would be the equation for the normal distribution, which depends only on μ and σ^2. Eliason (1993) provides a simple worked example.

The ML estimator can be determined graphically by simply trying different values of μ and seeing which one maximizes the log-likelihood function (Figure 2.3). This is very tedious, however, and it is easier (and more accurate) to use some simple calculus to determine the value of μ that maximizes the likelihood function. ML estimators sometimes have exact arithmetical solutions, such as when estimating means or parameters for linear models (Chapters 8–12). In contrast, when analyzing some non-normal distributions, ML estimators need to be calculated using complex iterative algorithms (Chapters 13 and 14).

It is important to realize that a likelihood is not the same as a probability and the likelihood function is not a probability distribution (Barnett 1999, Hilborn & Mangel 1997). In a probability distribution for a random variable, the parameter is considered fixed and the data are the unknown variable(s). In a likelihood function, the data are considered fixed and it is the parameter that varies across all possible values. However, the likelihood of the data given a particular parameter value is related to the probability of obtaining the data assuming this particular parameter value (Hilborn & Mangel 1997).

2.4.2 Ordinary least squares (OLS)

Another general approach to estimating parameters is by ordinary least squares (OLS). The least squares estimator for a given parameter is the one that minimizes the sum of the squared differences between each value in a sample and the parameter, i.e. minimizes the following function:

$$\sum_{i=1}^{n} [y_i - f(\theta)]^2 \qquad (2.12)$$

The OLS estimator of μ for a given sample is the value of μ which minimises the sum of squared differences between each value in the sample and the estimate of μ (i.e. $\Sigma(y_i - \bar{y})^2$). OLS estimators are usually more straightforward to calculate than ML estimators, always having exact arithmetical solutions. The major application of OLS estimation is when we are estimating parameters of linear models (Chapter 5 onwards), where Equation 2.12 represents the sum of squared differences between observed values and those predicted by the model.

2.4.3 ML vs OLS estimation

Maximum likelihood and ordinary least squares are not the only methods for estimating population parameters (see Barnett 1999) but they are the most commonly used for the analyses we will discuss in this book. Point and interval estimation using ML relies on distributional assumptions, i.e. we need to specify a probability distribution for our variable or for the error terms from our statistical model (see Chapter 5 onwards). When these assumptions are met, ML estimators are generally unbiased, for reasonable sample sizes, and they have minimum variance (i.e., they are precise estimators) compared to other estimators. In contrast, OLS point estimates require no distributional assumptions, and OLS estimators are also generally unbiased and have minimum variance. However, for interval estimation and hypothesis testing, OLS estimators have quite restrictive distributional assumptions related to normality and patterns of variance.

For most common population parameters (e.g. μ), the ML and OLS estimators are the same when the assumptions of OLS are met. The exception is σ^2 (the population variance) for which the ML estimator (which uses n in the denominator) is slightly biased, although the bias is trivial if the sample size is reasonably large (Neter *et al.* 1996). In balanced linear models (linear regression and ANOVA) for which the assumptions hold (see Chapter 5 onwards), ML and OLS estimators of regression slopes and/or factor effects are identical. However, OLS is inappropriate for some common models where the response variable(s) or the residuals are not distributed normally, e.g. binary and more general categorical data. Therefore, generalized linear modeling (GLMs such as logistic regression and log-linear models; Chapter 13) and nonlinear modeling (Chapter 6) are based around ML estimation.

2.5 Resampling methods for estimation

The methods described above for calculating standard errors for a statistic and confidence intervals for a parameter rely on knowing two properties of the statistic (Dixon 1993).

- The sampling distribution of the statistic, usually assumed to be normal, i.e. the Central Limit Theorem holds.
- The exact formula for the standard error (i.e. the standard deviation of the statistic).

These conditions hold for a statistic like the sample mean but do not obviously extend to other statistics like the median (Efron & Gong 1983). In biology, we would occasionally like to estimate the population values of many measurements for which the sampling distributions and variances are unknown. These include ecological indices such as the intrinsic rate of increase (r) and dissimilarity coefficients (Dixon 1993) and statistics from unusual types of analyses, such as the intercept of a smoothing function (see Chapter 5; Efron & Tibshirani 1991). To measure the precision (i.e. standard errors and confidence intervals) of these types of statistics we must rely on alternative, computer-intensive resampling methods. The two approaches described below are based on the same principle: in the absence of other information, the best guess for the distribution of the population is the observations we have in our sample. The methods estimate the standard error of a statistic and confidence intervals for a parameter by resampling from the original sample.

Good introductions to these methods include Crowley (1992), Dixon (1993), Manly (1997) and Robertson (1991), and Efron & Tibshirani (1991) suggest useful general applications. These resampling methods can also be used for hypothesis testing (Chapter 3).

2.5.1 Bootstrap

The bootstrap estimator was developed by Efron (1982). The sampling distribution of the statistic is determined empirically by randomly resampling (using a random number generator to choose the observations; see Robertson 1991), with replacement, from the original sample, usually with the same original sample size. Because sampling is with replacement, the same observation can obviously be resampled so the bootstrap samples will be different from each other. The desired statistic can be determined from each bootstrapped sample and the sampling distribution of each

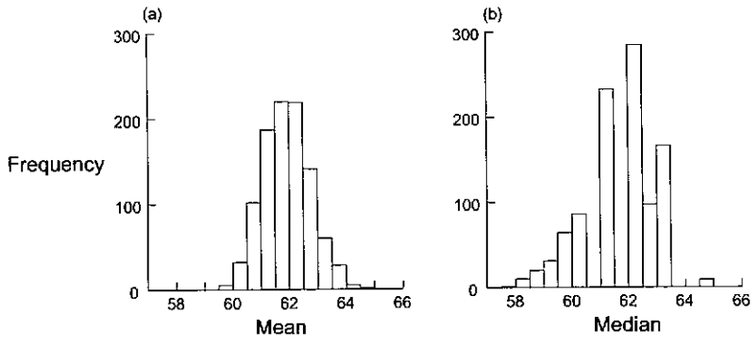

Figure 2.4 Frequency distributions of (a) bootstrap means and (b) bootstrap medians, based on 1000 bootstrap samples ($n = 39$) of SO_4^{2-} for 39 sites from forested watersheds in the Catskill Mountains in New York State (data from Lovett *et al.* 2000).

statistic determined. The bootstrap estimate of the parameter is simply the mean of the statistics from the bootstrapped samples. The standard deviation of the bootstrap estimate (i.e. the standard error of the statistic) is simply the standard deviation of the statistics from the bootstrapped samples (see Figure 2.4).

Techniques like the bootstrap can be used to measure the bias in an estimator, the difference between the actual population parameter and the expected value (mean) of the estimator. The bootstrap estimate of bias is simply the difference between the mean of the bootstrap statistics and the statistic calculated from the original sample (which is an estimator of the expected value of the statistic); see Robertson (1991).

Confidence intervals for the unknown population parameter can also be calculated based on the bootstrap samples. There are at least three methods (Dixon 1993, Efron & Gong 1983, Robertson 1991). First is the percentile method, where confidence intervals are calculated directly from the frequency distribution of bootstrap statistics. For example, we would arrange the 1000 bootstrap statistics in ascending order. Based on 1000 bootstrap samples, the lower limit of the 95% confidence interval would be the 25th value and the upper limit of the 95% confidence interval would be the 975th value; 950 values (95% of the bootstrap estimates) would fall between these values. Adjustments can easily be made for other confidence intervals, e.g. 5th and 995th value for a 99% confidence interval.

Unfortunately, the distribution of bootstrap statistics is often skewed, especially for statistics other than the mean. The confidence intervals calculated using the percentile method will not be symmetrical around the bootstrap estimate of the parameter, so the confidence intervals are biased. The other two methods for calculating bootstrap confidence intervals correct for this bias.

The bias-corrected method first works out the percentage of bootstrap samples with statistics lower than the bootstrap estimate. This is transformed to its equivalent value from the inverse cumulative normal distribution (z_0) and this value used to modify the percentiles used for the lower and upper limits of the confidence interval:

$$95\% \text{ percentiles} = \Phi(2z_0 \pm 1.96) \qquad (2.13)$$

where Φ is the normal cumulative distribution function. So we determine the percentiles for the values $(2z_0 + 1.96)$ and $(2z_0 - 1.96)$ from the normal cumulative distribution function and use these as the percentiles for our confidence interval. A worked example is provided in Box 2.2.

The third method, the accelerated bootstrap, further corrects for bias based on a measure of the influence each bootstrap statistic has on the final estimate. Dixon (1993) provides a readable explanation.

2.5.2 Jackknife

The jackknife is an historically earlier alternative to the bootstrap for calculating standard errors that is less computer intensive. The statistic is calculated from the full sample of n observations (call it θ^*), then from the sample with first data point removed (θ^*_{-1}), then from the sample with second data point removed (θ^*_{-2}) etc. Pseudovalues for each observation in the original sample are calculated as:

$$\tilde{\theta}_i = n\theta^* - (n-1)\theta^*_{-i} \qquad (2.14)$$

where θ^*_{-i} is the statistic calculated from the sample with observation i omitted. Each pseudovalue is simply a combination of two estimates of the statistic, one based on the whole sample and one based on the removal of a particular observation.

The jackknife estimate of the parameter is simply the mean of the pseudovalues ($\bar{\theta}$). The standard deviation of the jackknife estimate (the standard error of the estimate) is:

$$\sqrt{\frac{n-1}{n}\sum(\theta^*_{-i} - \bar{\theta})^2} \qquad (2.15)$$

Note that we have to assume that the pseudovalues are independent of each other for these calculations (Crowley 1992, Roberston 1991), whereas in reality they are not. The jackknife is not usually used for confidence intervals because so few samples are available if the original sample size was small (Dixon 1993). However, Crowley (1992) and Robertson (1991) suggested that if normality of the pseudovalues could be assumed, then confidence intervals could be calculated as usual (using the t distribution because of the small number of estimates).

2.6 | Bayesian inference – estimation

The classical approach to point and interval estimation might be considered to have two limitations. First, only the observed sample data contribute to our estimate of the population parameter. Any previous information we have on the likely value of the parameter cannot easily be considered when determining our estimate, although our knowledge of the population from which we are sampling will influence the design of our sampling program (Chapter 7). Second, the interval estimate we have obtained has a frequentist interpretation – a certain percentage of confidence intervals from repeated sampling will contain the fixed population parameter. The Bayesian approach to estimating parameters removes these limitations by formally incorporating our prior knowledge, as degrees-of-belief (Chapter 1), about the value of the parameter and by producing a probability statement about the parameter, e.g. there is a 95% probability that μ lies within a certain interval.

2.6.1 Bayesian estimation

To estimate parameters in a Bayesian framework, we need to make two major adjustments to the way we think about parameters and probabilities. First, we now consider the parameter to be a random variable that can take a range of possible values, each with different probabilities or degrees-of-belief of being true (Barnett 1999). This contrasts with the classical approach where the parameter was considered a fixed, but unknown, quantity. Dennis (1996), however, described the parameter being sought as an unknown variable rather than a random variable and the prior and posterior distributions represent the probabilities that this unknown parameter might take different values. Second, we must abandon our frequentist view of probability. Our interest is now only in the sample data we have, not in some long run hypothetical set of identical experiments (or samples). In Bayesian methods, probabilities can incorporate subjective degrees-of-belief (Chapter 1), although such opinions can still be quantified using probability distributions.

The basic logic of Bayesian inference for estimating a parameter is:

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})} \qquad (2.16)$$

where

θ is the population parameter to be estimated and is regarded as a random variable,

$P(\theta)$ is the "unconditional" prior probability of θ, expressed as a probability distribution summarizing our prior views about the probability of θ taking different values,

$P(\text{data}|\theta)$ is the likelihood of observing the sample data for different values of θ, expressed as a likelihood function (Section 2.4.1),

$P(\text{data})$ is the expected value (mean) of the likelihood function; this standardization means that the area under the posterior probability distribution equals one, and

$P(\theta|\text{data})$ is the posterior probability of θ conditional on the data being observed, expressed a probability distribution summarizing the probability of θ taking different values by combining the prior probability distribution and the likelihood function.

Equation 2.16 can be re-expressed more simply as:

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability} \quad (2.17)$$

because the denominator in Equation 2.15, $P(\text{data})$, is a normalizing constant, the mean of the likelihood function (Ellison 1996).

2.6.2 Prior knowledge and probability

Prior probability distributions measure the relative "strength of belief" in possible values of the parameter (Dennis 1996) and can be of two forms (Barnett 1999).

1. Prior ignorance or only vague prior knowledge, where we have little or no previous information to suggest what value the parameter might take. While some Bayesians might argue that scientists will always have some prior information, and that we will never be in a position of complete ignorance, prior ignorance is a conservative approach and helps overcome the criticism of Bayesian statistics that subjectively determined prior opinion can have too much influence on the inferential process. We can represent prior ignorance with a non-informative prior distribution, sometimes called a diffuse distribution because such a wide range of values of θ is considered possible. The most typical diffuse prior is a rectangular (uniform or flat) probability distribution, which says that each value of the parameter is equally likely.

One problem with uniform prior distributions is that they are improper, i.e. the probability distribution does not integrate to one and therefore the probability of any range of values might not be less than one. In practice, this is not a serious problem because improper priors can be combined with likelihoods to produce proper posterior distributions. When we use a non-informative prior, the posterior distribution of the parameter is directly proportional to the likelihood function anyway. The uniform prior distribution can be considered a reference prior, a class of priors designed to represent weak prior knowledge and let the data, and therefore the likelihood, dominate the posterior distribution.

2. Substantial prior knowledge or belief represented by an informative prior probability distribution such as a normal or beta distribution. The construction of these informative prior distributions is one of the most controversial aspects of Bayesian inference, especially if they are constructed from subjective opinion. Crome *et al.* (1996) illustrated one approach based on surveying a small group of people for the opinions about the effects of logging. Dennis (1996) and Mayo (1996) have respectively highlighted potential practical and philosophical issues associated with using subjective prior information.

2.6.3 Likelihood function

The likelihood function $P(\text{data}|\theta)$, standardized by the expected value (mean) of likelihood function $[P(\text{data})]$, is how the sample data enter Bayesian calculations. Note that the likelihood function is not strictly a probability distribution (Section 2.4.1), although we refer to it as the probability of observing the data for different values of the parameter. If we assume that our variable is normally distributed and the parameter of interest is the mean, the standardized likelihood function is a normal distribution with a mean equal to the mean of the sample data and a variance equal to the squared standard error of the mean of the sample data (Box & Tiao 1973, Ellison 1996).

2.6.4 Posterior probability

All conclusions from Bayesian inference are based on the posterior probability distribution of the parameter. This posterior distribution represents our prior probability distribution modified by the likelihood function. The sample data only enter Bayesian inference through the likelihood function. Bayesian inference is usually based on the shape of the posterior distribution, particularly the range of values over which most of the probability mass occurs. The best estimate of the parameter is determined from the mean of the posterior distribution, or sometimes the median or mode if we have a non-symmetrical posterior.

If we consider estimating a parameter (θ) with a normal prior distribution, then the mean of the normal posterior distribution of θ is (Box & Tiao 1973, Ellison 1996):

$$\bar{\theta} = \frac{1}{w_0 + w_1}(w_0\bar{\theta}_0 + w_1\bar{y}) \quad (2.18)$$

where $\bar{\theta}_0$ is the mean of the prior distribution, $\bar{y}$ is the mean of the likelihood function (i.e. sample mean from data), w_0 is the reciprocal of the estimate of the prior variance σ_0^2 $(1/s_0^2)$, w_1 is the reciprocal of the sample variance times the sample size (n/s^2) and n is the sample size. In other words, the posterior mean is a weighted average of the prior mean and the sample mean (Berry 1996). This posterior mean $\bar{\theta}$ is our estimate of θ, the parameter of interest.

The variance of the posterior distribution equals:

$$\bar{\sigma}^2 = \frac{1}{w_0 + w_1} \quad (2.19)$$

Note that with a non-informative, flat, prior the posterior distribution is determined entirely by the sample data and the likelihood function. The mean of the posterior then is $\bar{y}$ (the mean of the sample data) and the variance is s^2/n (the variance of the sample data divided by the sample size).

The Bayesian analogues of frequentist confidence intervals are termed Bayesian credible or probability intervals. They are also called highest density or probability regions because any value in the region or interval has a higher probability of occurring than any value outside. If we have a normal posterior distribution for a parameter, Bayesian credible intervals for this parameter are:

$$P\{\bar{\theta} - 2\sqrt{D} \leq \theta \leq \bar{\theta} + 2\sqrt{D}\} = 0.95 \quad (2.20)$$

where $D = \bar{\sigma}^2$, the variance of the posterior distribution (Ellison 1996). Alternatively, the usual methods based on the t distribution can be used (Winkler 1993). Note that because the parameter is considered a random variable in Bayesian inference, the interval in Equation 2.20 is telling us directly that there is a 95% probability that the value of the parameter falls within this range, based on the sample data. With a non-informative (flat) prior distribution, the Bayesian confidence interval will be the same as the classical, frequentist, confidence interval and Edwards (1996) argued that the difference in interpretation is somewhat semantic. He recommended simply reporting the interval and letting the reader interpret it as required. If we have a more informative prior distribution (i.e. we knew that some values of θ were more likely than others), then the Bayesian credible interval would be shorter than the classical confidence interval.

2.6.5 Examples

We provide a very simple example of Bayesian estimation in Box 2.4, based on the data from Lovett *et al.* (2000) on the chemistry of forested watersheds. Another biological example of Bayesian estimation is the work of Carpenter (1990). He compared eight different models for flux of pesticides through a pond ecosystem. Each model was given an equal prior probability (0.125), data were collected from an experiment using radioactively labeled pesticide and likelihoods were determined for each model from the residuals after each model was fitted using OLS (see Chapter 2). He found that only one of the models had a posterior probability greater than 0.1 (actually it was 0.97, suggesting it was a very likely outcome).

2.6.6 Other comments

We would like to finish with some comments. First, normal distributions are commonly used for both prior and posterior distributions and likelihood functions for the same reasons as for classical estimation, especially when dealing with means. Other distributions can be used. For example, Crome *et al.* (1996) used a mixture of log-normal distributions for an informative prior (see also Winkler 1993) and the beta distribution is commonly used as a prior for binomially distributed parameters.

Second, the data generally are much more influential over the posterior distribution than the prior, except when sample sizes, and/or the variance of the prior, are very small. Carpenter (1990) discussed Bayesian analysis in the context of large-scale perturbation experiments in ecology and he also argued that prior probabilities had far less impact than the observed data on the outcome of the analysis and implied that the choice of prior probabilities was not crucial. However, Edwards (1996) noted that if the prior standard deviation is very small, then differences in the prior mean could have marked effects on

Box 2.4 | Worked example of Bayesian estimation: chemistry of forested watersheds

To illustrate the Bayesian approach to estimation, we will revisit the earlier example of estimating the mean concentration of SO_4^{2-} in first and second order stream sites in the Catskill Mountains in New York State based on a sample of 39 sites (Lovett *et al.* 2000). Now we will consider the mean concentration of SO_4^{2-} a random variable, or at least an unknown variable (Dennis 1996), and also make use of prior information about this mean, i.e. we will estimate our mean from a Bayesian perspective. For comparison, we will also investigate the effect of more substantial prior knowledge, in the form of a less variable prior probability distribution. We will follow the procedure for Bayesian estimation from Box & Tiao (1973; see also Berry 1996 and Ellison 1996).

1. Using whatever information is available (including subjective assessment; see Crome *et al.* 1996), specify a prior probability distribution for Y. Note that initial estimates of the parameters of this distribution will need to be specified; a normal prior requires an initial estimate of the mean and variance. Imagine we had sampled the central Catskill Mountains at a previous time so we had some previous data that we could use to set up a prior distribution. We assumed the prior distribution of the concentration of SO_4^{2-} was normal and we used the mean and the variance of the previous sample as the parameters of the prior distribution. The prior distribution could also be a non-informative (flat) one if no such previous information was available.

2. Collect a sample to provide an estimate of the parameter and its variance. In our example, we had a sample of concentration of SO_4^{2-} from 39 streams and determined the sample mean and variance.

3. Determine the standardized likelihood function, which in this example is a normal distribution with a mean equal to the mean of the sample data and a variance equal to the squared standard error of the mean of the sample data.

4. Determine the posterior probability distribution for the mean concentration of SO_4^{2-}, which will be a normal distribution because we used a normal prior and likelihood function. The mean of this posterior distribution (Equation 2.18) is our estimate of population mean concentration of SO_4^{2-} and we can determine credible intervals for this mean (Equation 2.20).

High variance prior distribution

Prior mean = 50.00, prior variance = 44.00.
Sample mean = 61.92, sample variance = 27.47, n = 39.
Using Equations 2.18, 2.19 and 2.20, substituting sample estimates where appropriate:
$w_0 = 0.023$
$w_1 = 1.419$
Posterior mean = 61.73, posterior variance = 0.69, 95% Bayesian probability interval = 60.06 to 62.57.
Note that the posterior distribution has almost the same estimated mean as the sample, so the posterior is determined almost entirely by the sample data.

Low variance prior distribution

If we make our prior estimate of the mean much more precise:
Prior mean = 50.00, prior variance = 10.00.
Sample mean = 61.92, sample variance = 27.47, n = 39.
$w_0 = 0.100$
$w_1 = 1.419$
Posterior mean = 61.14, posterior variance = 0.66, 95% Bayesian probability interval = 59.51 to 62.76.

Now the prior distribution has a greater influence on the posterior than previously, with the posterior mean more than half one unit lower. In fact, the more different the prior mean is from the sample mean, and the more precise our estimate of the prior mean is, i.e. the lower the prior variance, the more the prior will influence the posterior relative to the data.

Note that if we assume a flat prior, the posterior mean is just the mean of the data (61.92).

the posterior mean, irrespective of the data. He described this as "editorial", where the results of the analysis are mainly opinion.

Third, if a non-informative prior (like a rectangular distribution) is used, and we assume the data are from a normally distributed population, then the posterior distribution will be a normal (or t) distribution just like in classical estimation, i.e. using a flat prior will result in the same estimates as classical statistics. For example, if we wish to use Bayesian methods to estimate μ, and we use a rectangular prior distribution, then the posterior distribution will turn out to be a normal distribution (if σ is known) or a t distribution (if σ is unknown and estimated from s, which means we need a prior distribution for s as well).

Finally, we have provided only a very brief introduction to Bayesian methods for estimation and illustrated the principle with a simple example. For more complex models with two or more parameters, calculating the posterior distribution is difficult. Recent advances in this area use various sampling algorithms (e.g. Hastings-Metropolis Gibbs sampler) as part of Markov chain Monte Carlo methods. These techniques are beyond the scope of this book – Barnett (1999) and Gelman *et al.* (1995) provide an introduction although the details are not for the mathematically challenged. The important point is that once we get beyond simple estimation problems, Bayesian methods can involve considerable statistical complexity.

Other pros and cons related to Bayesian inference, particularly in comparison with classical frequentist inference, will be considered in Chapter 3 in the context of testing hypotheses.