CHAPTER 4

Framing and Testing Hypotheses

Hypotheses are potential explanations that can account for our observations of the external world. They usually describe cause-and-effect relationships between a proposed mechanism or process (the cause) and our observations (the effect). Observations are data—what we see or measure in the real world. Our goal in undertaking a scientific study is to understand the cause(s) of observable phenomena. Collecting observations is a means to that end: we accumulate different kinds of observations and use them to distinguish among different possible causes. Some scientists and statisticians distinguish between observations made during manipulative experiments and those made during **observational** or **correlative studies**. However, in most cases, the statistical treatment of such data is identical. The distinction lies in the confidence we can place in **inferences**[1] drawn from those studies. Well-designed manipulative experiments allow us to be confident in our inferences; we have less confidence in data from poorly designed experiments, or from studies in which we were unable to directly manipulate variables.

If observations are the "what" of science, hypotheses are the "how." Whereas observations are taken from the real world, hypotheses need not be. Although our observations may suggest hypotheses, hypotheses can also come from the existing body of scientific literature, from the predictions of theoretical models, and from our own intuition and reasoning. However, not all descriptions of cause-and-effect relationships constitute valid scientific hypotheses. A scientific hypothesis must be testable: in other words, there should be some set of additional observations or experimental results that we could collect that would cause

[1] In logic, an *inference* is a conclusion that is derived from premises. Scientists make inferences (draw conclusions) about causes based on their data. These conclusions may be suggested, or implied, by the data. But remember that it is the scientist who infers, and the data that imply.

us to modify, reject, or discard our working hypothesis.[2] Metaphysical hypotheses, including the activities of omnipotent gods, do not qualify as scientific hypotheses because these explanations are taken on faith, and there are no observations that would cause a believer to reject these hypotheses.[3]

In addition to being testable, a good scientific hypothesis should generate novel predictions. These predictions can then be tested by collecting additional observations. However, the same set of observations may be predicted by more than one hypothesis. Although hypotheses are chosen to account for our initial observations, a good scientific hypothesis also should provide a unique set of predictions that do not emerge from other explanations. By focusing on these unique predictions, we can collect more quickly the critical data that will discriminate among the alternatives.

Scientific Methods

The "scientific method" is the technique used to decide among hypotheses on the basis of observations and predictions. Most textbooks present only a single scientific method, but practicing scientists actually use several methods in their work.

---

[2] A scientific *hypothesis* refers to a particular mechanism or cause-and-effect relationship; a scientific *theory* is much broader and more synthetic. In its early stages, not all elements of a scientific theory may be fully articulated, so that explicit hypotheses initially may not be possible. For example, Darwin's theory of natural selection required a mechanism of inheritance that conserved traits from one generation to the next while still preserving variation among individuals. Darwin did not have an explanation for inheritance, and he discussed this weakness of his theory in *The Origin of Species* (1859). Darwin did not know that precisely such a mechanism had in fact been discovered by Gregor Mendel in his experimental studies (ca. 1856) of inheritance in pea plants. Ironically, Darwin's grandfather, Erasmus Darwin, had published work on inheritance two generations earlier, in his *Zoonomia, or the Laws of Organic Life* (1794–1796). However, Erasmus Darwin used snapdragons as his experimental organism, whereas Mendel used pea plants. Inheritance of flower color is simpler in pea plants than in snapdragons, and Mendel was able to recognize the particulate nature of genes, whereas Erasmus Darwin could not.

[3] Although many philosophies have attempted to bridge the gap between science and religion, the contradiction between reason and faith is a critical fault line separating the two. The early Christian philosopher Tertullian (~155–222 AD) seized upon this contradiction and asserted "*Credo quai absurdum est*" ("I believe because it is absurd"). In Tertullian's view, that the son of God died is to be believed because it is contradictory; and that he rose from the grave has certitude because it is impossible (Reese 1980).

Deduction and Induction

Deduction and induction are two important modes of scientific reasoning, and both involve drawing inferences from data or models. **Deduction** proceeds from the general case to the specific case. The following set of statements provides an example of classic deduction:

1. All of the ants in the Harvard Forest belong to the genus *Myrmica*.
2. I sampled this particular ant in the Harvard Forest.
3. This particular ant is in the genus *Myrmica*.

Statements 1 and 2 are usually referred to as the *major and minor premises*, and statement 3 is the *conclusion*. The set of three statements is called a **syllogism**, an important logical structure developed by Aristotle. Notice that the sequence of the syllogism proceeds from the general case (all of the ants in the Harvard Forest) to the specific case (the particular ant that was sampled).

In contrast, **induction** proceeds from the specific case to the general case:[4]

1. All 25 of these ants are in the genus *Myrmica*.
2. All 25 of these ants were collected in the Harvard Forest.
3. All of the ants in the Harvard Forest are in the genus *Myrmica*.

Sir Francis Bacon

[4] The champion of the inductive method was Sir Francis Bacon (1561–1626), a major legal, philosophical, and political figure in Elizabethan England. He was a prominent member of parliament, and was knighted in 1603. Among scholars who question the authorship of Shakespeare's works (the so-called anti-Stratfordians), some believe Bacon was the true author of Shakespeare's plays, but the evidence isn't very compelling. Bacon's most important scientific writing is the *Novum organum* (1620), in which he urged the use of induction and empiricism as a way of knowing the world. This was an important philosophical break with the past, in which explorations of "natural philosophy" involved excessive reliance on deduction and on published authority (particularly the works of Aristotle). Bacon's inductive method paved the way for the great scientific breakthroughs by Galileo and Newton in the Age of Reason. Near the end of his life, Bacon's political fortunes took a turn for the worse; in 1621 he was convicted of accepting bribes and was removed from office. Bacon's devotion to empiricism eventually did him in. Attempting to test the hypothesis that freezing slows the putrefaction of flesh, Bacon ventured out in the cold during the winter of 1626 to stuff a chicken with snow. He became badly chilled and died a few days later at the age of 65.

Some philosophers define deduction as certain inference and induction as probable inference. These definitions certainly fit our example of ants collected in the Harvard Forest. In the first set of statements (deduction), the conclusion must be logically true if the first two premises are true. But in the second case (induction), although the conclusion is *likely* to be true, it may be false; our confidence will increase with the size of our sample, as is always the case in statistical inference. Statistics, by its very nature, is an inductive process: we are always trying to draw general conclusions based on a specific, limited sample.

Both induction and deduction are used in all models of scientific reasoning, but they receive different emphases. Even using the inductive method, we probably will use deduction to derive specific predictions from the general hypothesis in each turn of the cycle.

Any scientific inquiry begins with an observation that we are trying to explain. The inductive method takes this observation and develops a single hypothesis to explain it. Bacon himself emphasized the importance of using the data to suggest the hypothesis, rather than relying on conventional wisdom, accepted authority, or abstract philosophical theory. Once the hypothesis is formulated, it generates—through deduction—further predictions. These predictions are then tested by collecting additional observations. If the new observations match the predictions, the hypothesis is supported. If not, the hypothesis must be modified to take into account both the original observation and the new observations. This cycle of hypothesis–prediction–observation is repeatedly traversed. After each cycle, the modified hypothesis should come closer to the truth[5] (Figure 4.1).

Two advantages of the inductive method are (1) it emphasizes the close link between data and theory; and (2) it explicitly builds and modifies hypotheses based on previous knowledge. The inductive method is *confirmatory* in that we

[5] Ecologists and environmental scientists rely on induction when they use statistical software to fit non-linear (curvy) functions to data (see Chapter 9). The software requires that you specify not only the equation to be fit, but also a set of initial values for the unknown parameters. These initial values need to be "close to" the actual values because the algorithms are *local* estimators (i.e., they solve for local minima or maxima in the function). Thus, if the initial estimates are "far away" from the actual values of the function, the curve-fitting routines may either fail to converge on a solution, or will converge on a non-sensical one. Plotting the fitted curve along with the data is a good safeguard to confirm that the curve derived from the estimated parameters actually fits the original data.

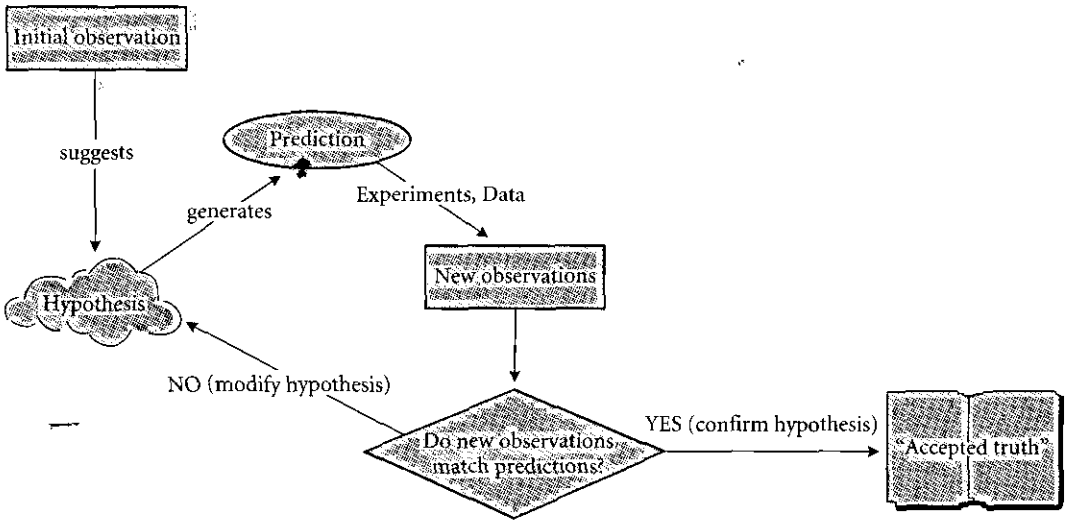
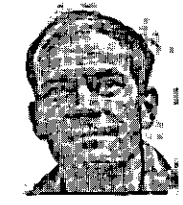


Figure 4.1 The inductive method. The cycle of hypothesis, prediction, and observation is repeatedly traversed. Hypothesis confirmation represents the theoretical endpoint of the process. Compare the inductive method to the hypothetico-deductive method (Figure 4.4), in which multiple working hypotheses are proposed and emphasis is placed on falsification rather than verification.

seek data that support the hypothesis, and then we modify the hypothesis to conform with the accumulating data.[6]

There are also several disadvantages of the inductive method. Perhaps the most serious is that the inductive method considers only a single starting hypothesis; other hypotheses are only considered later, in response to additional data and observations. If we "get off on the wrong foot" and begin exploring an incorrect hypothesis, it may take a long time to arrive at the correct answer


Robert H. MacArthur

[6] The community ecologist Robert H. MacArthur (1930–1972) once wrote that the group of researchers interested in making ecology a science "arranges ecological data as examples testing the proposed theories and spends most of its time patching up the theories to account for as many of the data as possible." (MacArthur 1962). This quote characterizes much of the early theoretical work in community ecology. Later, theoretical ecology developed as a discipline in its own right, and some interesting lines of research blossomed without any reference to data or the real world. Ecologists disagree about whether such a large body of purely theoretical work has been good or bad for our science (Pielou 1981; Caswell 1988).

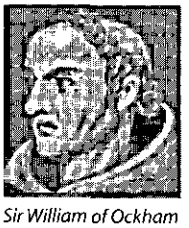
through induction. In some cases we may never get there at all. In addition, the inductive method may encourage scientists to champion pet hypotheses and perhaps hang on to them long after they should have been discarded or radically modified (Loehle 1987). And finally, the inductive method—at least Bacon's view of it—derives theory exclusively from empirical observations. However, many important theoretical insights have come from theoretical modeling, abstract reasoning, and plain old intuition. Important hypotheses in all sciences have often emerged well in advance of the critical data that are needed to test them.[7]

Modern-Day Induction: Bayesian Inference

The **null hypothesis** is the starting point of a scientific investigation. A null hypothesis tries to account for patterns in the data in the simplest way possible, which often means initially attributing variation in the data to randomness or measurement error. If that simple null hypothesis can be rejected, we can move on to entertain more complex hypotheses.[8] Because the inductive method begins with an observation that suggests an hypothesis, how do we generate an appropriate null hypothesis? Bayesian inference represents a modern, updated version of the inductive method. The principals of Bayesian inference can be illustrated with a simple example.

The photosynthetic response of leaves to increases in light intensity is a well-studied problem. Imagine an experiment in which we grow 15 mangrove seedlings, each under a different light intensity (expressed as photosynthetic photon flux density, or PPFD, in μmol photons per m^2 of leaf tissue exposed to

[7] For example, in 1931 the Austrian physicist Wolfgang Pauli (1900–1958) hypothesized the existence of the neutrino, an electrically neutral particle with negligible mass, to account for apparent inconsistencies in the conservation of energy during radioactive decay. Empirical confirmation of the existence of neutrino did not come until 1956.



Sir William of Ockham

[8] The preference for simple hypotheses over complex ones has a long history in science. Sir William of Ockham's (1290–1349) Principle of Parsimony states that "[Entities] are not to be multiplied beyond necessity." Ockham believed that unnecessarily complex hypotheses were vain and insulting to God. The Principle of Parsimony is sometimes known as Ockham's Razor, the razor shearing away unnecessary complexity. Ockham lived an interesting life. He was educated at Oxford and was a member of the Franciscan order. He was charged with heresy for some of the writing in his Master's thesis. The charge was eventually dropped, but when Pope John XXII challenged the Franciscan doctrine of apostolic poverty, Ockham was excommunicated and fled to Bavaria. Ockham died in 1349, probably a victim of the bubonic plague epidemic.

light each second) and measure the photosynthetic response of each plant (expressed as μmoles of CO_2 fixed per m^2 of leaf tissue exposed to light each second). We then plot the data with light intensity on the *x*-axis (the *predictor variable*) and photosynthetic rate on the *y*-axis (the *response variable*). Each point represents a different leaf for which we have recorded these two numbers.

In the absence of any information about the relationship between light and photosynthetic rates, the simplest null hypothesis is that there is no relationship between these two variables (Figure 4.2). If we fit a line to this null hypothesis, the slope of the line would equal 0. If we collected data and found some other relationship between light availability and photosynthetic rate, we would then use those data to modify our hypothesis, following the inductive method.

But is it really necessary to frame the null hypothesis as if you had no information at all? Using just a bit of knowledge about plant physiology, we can formulate a more realistic initial hypothesis. Specifically, we expect there to be some

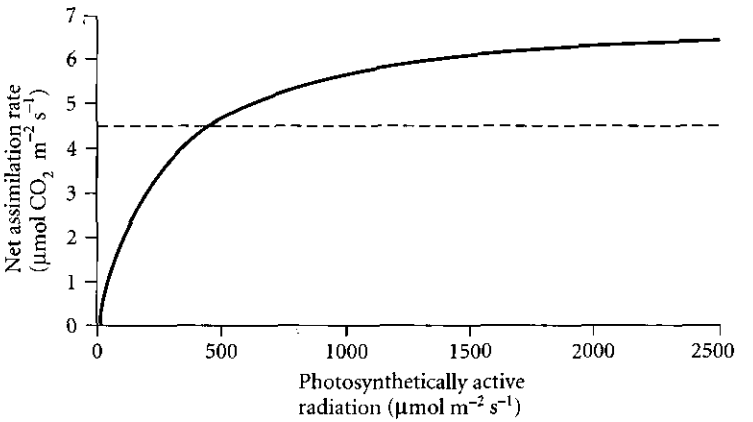


Figure 4.2 Two null hypotheses for the relationship between light intensity (measured as photosynthetically active radiation) and photosynthetic rate (measured as net assimilation rate) in plants. The simplest null hypothesis is that there is no association between the two variables (dashed line). This null hypothesis is the starting point for a hypothetico-deductive approach that assumes no prior knowledge about the relationship between the variables and is the basis for a standard linear regression model (Chapter 9). In contrast, the green curve represents a Bayesian approach of bringing prior knowledge to create an informed null hypothesis. In this case, the "prior knowledge" is of plant physiology and photosynthesis. We expect that the assimilation rate will rise rapidly at first as light intensity increases, but then reach an asymptote or saturation level. Such a relationship can be described by a Michaelis-Menten equation [$Y = kX/(D + X)$], which includes parameters for an asymptotic assimilation rate (k) and a half saturation constant (D) that controls the steepness of the curve. Bayesian methods can incorporate this type of prior information into the analysis.

maximum photosynthetic rate that the plant can achieve. Beyond this point, increases in light intensity will not yield additional photosynthate, because some other factor, such as water or nutrients, becomes limiting. Even if these factors were supplied and the plant were grown under optimal conditions, photosynthetic rates will still level out because there are inherent limitations in the rates of biochemical processes and electron transfers that occur during photosynthesis. (In fact, if we keep increasing light intensity, excessive light energy can damage plant tissues and reduce photosynthesis. But in our example, we limit the upper range of light intensities to those that the plant can tolerate.)

Thus, our informed null hypothesis is that the relationship between photosynthetic rate and light intensity should be non-linear, with an asymptote at high light intensities (see Figure 4.2). Real data could then be used to test the degree of support for this more realistic null hypothesis (Figure 4.3). To determine which null hypothesis to use, we also must ask what, precisely, is the point of the study? The simple null hypothesis (linear equation) is appropriate if we merely want to establish that a non-random relationship exists between light intensity and photosynthetic rate. The informed null hypothesis (Michaelis-Menten equation) is appropriate if we want to compare saturation curves among species or to test theoretical models that make quantitative predictions for the asymptote or half-saturation constant.

Figures 4.2 and 4.3 illustrate how a modern-day inductivist, or Bayesian statistician, generates an hypothesis. The Bayesian approach is to use prior knowledge or information to generate and test hypotheses. In this example, the prior knowledge was derived from plant physiology and the expected shape of the light saturation curve. However, prior knowledge might also be based

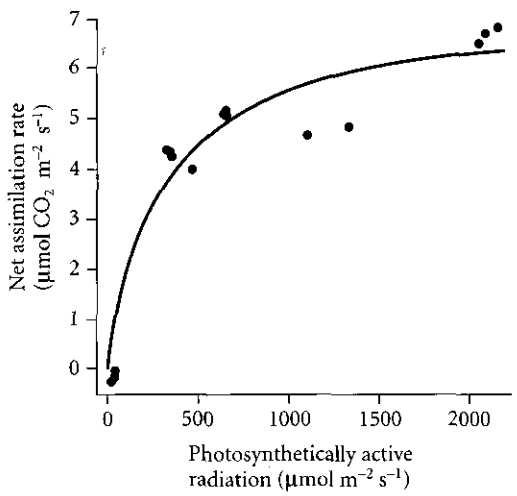


Figure 4.3 Relationship between light intensity and photosynthetic rate. The data are measurements of net assimilation rate and photosynthetically active radiation for $n = 15$ young sun leaves of the mangrove *Rhizophora mangle* in Belize (Farnsworth and Ellison 1996b). A Michaelis-Menten equation of the form $Y = kX/(D + X)$ was fit to the data. The parameter estimates ± 1 standard error are $k = 7.3 \pm 0.58$ and $D = 313 \pm 86.6$.

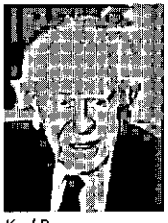
on the extensive base of published literature on light saturation curves (Björkman 1981; Lambers et al. 1998). If we had empirical parameter estimates from other studies, we could quantify our prior estimates of the threshold and asymptote values for light saturation. These estimates could then be used to further specify the initial hypothesis for fitting the asymptote value to our experimental data.

Use of prior knowledge in this way is different from Bacon's view of induction, which was based entirely on an individual's own experience. In a Baconian universe, if you had never studied plants before, you would have no direct evidence on the relationship between light and photosynthetic rate, and you would begin with a null hypothesis such as the flat line shown in Figure 4.2. This is actually the starting point for the hypothetico-deductive method presented in the next section.

The strict Baconian interpretation of induction is the basis of the fundamental critique of the Bayesian approach: that the prior knowledge used to develop the initial model is arbitrary and subjective, and may be biased by preconceived notions of the investigator. Thus, the hypothetico-deductive method is viewed by some as more "objective" and hence more "scientific." Bayesians counter this argument by asserting that the statistical null hypotheses and curve-fitting techniques used by hypothetico-deductivists are just as subjective; these methods only seem to be more objective because they are familiar and uncritically accepted. For a further discussion of these philosophical issues, see Ellison (1996, 2004) and Dennis (1996).

The Hypothetico-Deductive Method

The **hypothetico-deductive method** (Figure 4.4) developed from the works of Sir Isaac Newton and other seventeenth-century scientists and was championed by the philosopher of science Karl Popper.[9] Like the inductive method, the hypothetico-deductive method begins with an initial observation that we are trying to explain. However, rather than positing a single hypothesis and working



Karl Popper

[9] The Austrian philosopher of science Karl Popper (1902–1994) was the most articulate champion of the hypothetico-deductive method and falsifiability as the cornerstone of science. In *The Logic of Scientific Discovery* (1935), Popper argued that falsifiability is a more reliable criterion of truth than verifiability. In *The Open Society and Its Enemies* (1945), Popper defended democracy and criticized the totalitarian implications of induction and the political theories of Plato and Karl Marx.

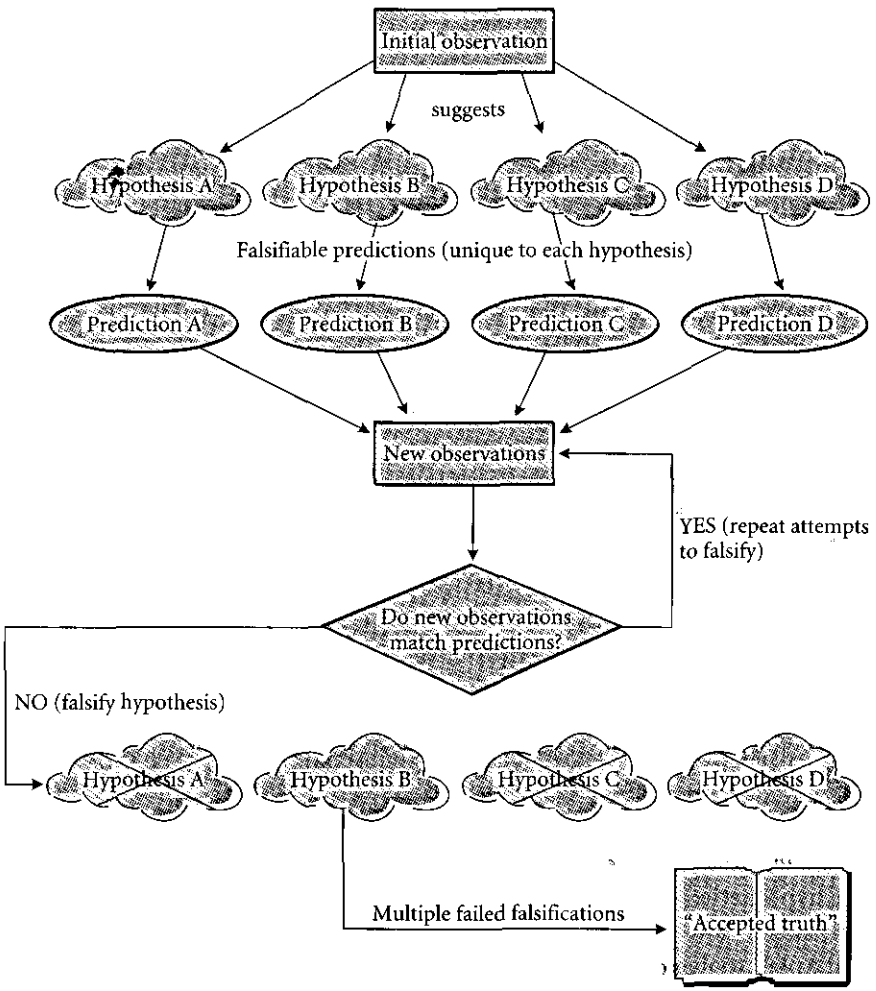


Figure 4.4 The hypothetico-deductive method. Multiple working hypotheses are proposed and their predictions tested with the goal of falsifying the incorrect hypotheses. The correct explanation is the one that stands up to repeated testing but fails to be falsified.

forward, the hypothetico-deductive method asks us to propose multiple, working hypotheses. All of these hypotheses account for the initial observation, but they each make additional unique predictions that can be tested by further experiments or observations. The goal of these tests is not to confirm, but to falsify, the hypotheses. Falsification eliminates some of the explanations, and the list is winnowed down to a smaller number of contenders. The cycle of predictions

and new observations is repeated. However, the hypothetico-deductive method never confirms an hypothesis; the accepted scientific explanation is the hypothesis that successfully withstands repeated attempts to falsify it.

The two advantages of the hypothetico-deductive method are: (1) it forces a consideration of multiple working hypotheses right from the start; and (2) it highlights the key predictive differences between them. In contrast to the inductive method, hypotheses do not have to be built up from the data, but can be developed independently or in parallel with data collection. The emphasis on falsification tends to produce simple, testable hypotheses, so that parsimonious explanations are considered first, and more complicated mechanisms only later.[10]

The disadvantages of the hypothetico-deductive method are that multiple working hypotheses may not always be available, particularly in the early stages of investigation. Even if multiple hypotheses are available, the method does not really work unless the "correct" hypothesis is among the alternatives. In contrast, the inductive method may begin with an incorrect hypothesis, but can reach the correct explanation through repeated modification of the original hypothesis, as informed by data collection. Another useful distinction is that the inductive method gains strength by comparing many datasets to a single hypothesis, whereas the hypothetico-deductive method is best for comparing a single dataset to multiple hypotheses. Finally, both the inductive method and hypothetico-deductive method place emphasis on a single correct hypothesis, making it difficult to evaluate cases in which multiple factors are at work. This is less of a problem with the inductive approach, because multiple explanations can be incorporated into more complex hypotheses.

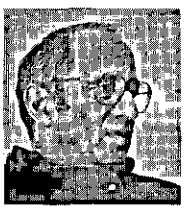
[10] The **logic tree** is a well-known variant of the hypothetico-deductive method that you may be familiar with from chemistry courses. The logic tree is a dichotomous decision tree in which different branches are followed depending on the results of experiments at each fork in the tree. The terminal branch tips of the tree represent the different hypotheses that are being tested. The logic tree also can be found in the familiar dichotomous taxonomic key for identifying to species unknown plants or animals: "If the animal has 3 pairs of walking legs, go to couplet *x*; if it has 4 or more pairs, go to couplet *y*." The logic tree is not always practical for complex ecological hypotheses; there may be too many branch points, and they may not all be dichotomous. However, it is always an excellent exercise to try and place your ideas and experiments in such a comprehensive framework. Platt (1964) champions this method and points to its spectacular success in molecular biology; the discovery of the helical structure of DNA is a classic example of the hypothetico-deductive method (Watson and Crick 1953).

Neither scientific method is the correct one, and some philosophers of science deny that either scenario really describes how science operates.[11] However, the hypothetico-deductive and inductive methods do characterize much science in the real world (as opposed to the abstract world of the philosophy of science). The reason for spending time on these models is to understand their relationship to statistical tests of an hypothesis.

Testing Statistical Hypotheses

Statistical Hypotheses versus Scientific Hypotheses

Using statistics to test hypotheses is only a small facet of the scientific method, but it consumes a disproportionate amount of our time and journal space. We use statistics to describe patterns in our data, and then we use statistical tests to decide whether the predictions of an hypothesis are supported or not. Establishing hypotheses, articulating their predictions, designing and executing valid experiments, and collecting, organizing, and summarizing the data all occur before we use statistical tests. We emphasize that accepting or rejecting a statistical hypothesis is quite distinct from accepting or rejecting a scientific hypothesis. The statistical null hypothesis is usually one of "no pattern," such as no difference between groups, or no relationship between two continuous variables. In contrast, the alternative hypothesis is that pattern exists. In other words, there



Thomas Kuhn

[11] No discussion of Popper and the hypothetico-deductive method would be complete without mention of Popper's philosophical nemesis, Thomas Kuhn (1922–1996). In *The Structure of Scientific Revolutions* (1962), Kuhn called into question the entire framework of hypothesis testing, and argued that it did not represent the way that science was done. Kuhn believed that science was done within the context of major **paradigms**, or research frameworks, and that the domain of these paradigms was implicitly adopted by each generation of scientists. The "puzzle-solving" activities of scientists constitute "ordinary science," in which empirical anomalies are reconciled with the existing paradigm. However, no paradigm can encompass all observations, and as anomalies accumulate, the paradigm becomes unwieldy. Eventually it collapses, and there is a scientific revolution in which an entirely new paradigm replaces the existing framework. Taking somewhat of an intermediate stance between Popper and Kuhn, the philosopher Imre Lakatos (1922–1974) thought that scientific research programs (SRPs) consisted of a core of central principles that generated a belt of surrounding hypotheses that make more specific predictions. The predictions of the hypotheses can be tested by the scientific method, but the core is not directly accessible (Lakatos 1978). Exchanges between Kuhn, Popper, Lakatos, and other philosophers of science can be read in Lakatos and Musgrave (1970). See also Horn (1986) for further discussion of these ideas.

are distinct differences in measured values between groups, or a clear relationship exists between two continuous variables. You must ask how such patterns relate to the scientific hypothesis you are testing.

For example, suppose you are evaluating the scientific hypothesis that waves scouring a rocky coast create empty space by removing competitively dominant invertebrate species. The open space can be colonized by competitively subordinate species that would otherwise be excluded. This hypothesis predicts that species diversity of marine invertebrates will change as a function of level of disturbance (Sousa 1979). You collect data on the number of species on disturbed and undisturbed rock surfaces. Using an appropriate statistical test, you find no difference in species richness in these two groups. In this case, you have *failed to reject* the statistical null hypothesis, and the pattern of the data *fail to support* one of the predictions of the disturbance hypothesis. Note, however, that absence of evidence is not evidence of absence; failure to reject a null hypothesis is not equivalent to accepting a null hypothesis (although it is often treated that way).

Here is a second example in which the statistical pattern is the same, but the scientific conclusion is different. The ideal free distribution is an hypothesis that predicts that organisms move between habitats and adjust their density so that they have the same mean fitness in different habitats (Fretwell and Lucas 1970). One testable prediction of this hypothesis is that the fitness of organisms in different habitats is similar, even though population density may differ. Suppose you measure population growth rate of birds (an important component of avian fitness) in forest and field habitats as a test of this prediction (Gill et al. 2001). As in the first example, you fail to reject the statistical null hypothesis, so that there is no evidence that growth rates differ among habitats. But in this case, *failing to reject* the statistical null hypothesis actually *supports* a prediction of the ideal free distribution.

Naturally, there are many additional observations and tests we would want to make to evaluate the disturbance hypothesis or the ideal free distribution. The point here is that the scientific and statistical hypotheses are distinct entities. In any study, you must determine whether supporting or refuting the statistical null hypothesis provides positive or negative evidence for the scientific hypothesis. Such a determination also influences profoundly how you set up your experimental study or observational sampling protocols. The distinction between the statistical null hypothesis and the scientific hypothesis is so important that we will return to it later in this chapter.

Statistical Significance and *P*-Values

It is nearly universal to report the results of a statistical test in order to assert the importance of patterns we observe in the data we collect. A typical assertion is: "The control and treatment groups differed significantly from one another

($P = 0.01$)." What, precisely, does "$P = 0.01$" mean, and how does it relate to the concepts of probability that we introduced in Chapters 1 and 2?

AN HYPOTHETICAL EXAMPLE: COMPARING MEANS A common assessment problem in environmental science is to determine whether or not human activities result in increased stress in animals. In vertebrates, stress can be measured as levels of the glucocorticoid hormones (GC) in the bloodstream or feces. For example, wolves that are not exposed to snowmobiles have 872.0 ng GC/g, whereas wolves exposed to snowmobiles have 1468.0 ng GC/g (Creel et al. 2002). Now, how do you decide whether this difference is large enough to be attributed to the presence of snowmobiles?[12]

Here is where you could conduct a conventional statistical test. Such tests can be very simple (such as the familiar *t*-test), or rather complex (such as tests for interaction terms in an analysis of variance). But all such statistical tests produce as their output a **test statistic**, which is just the numerical result of the test, and a **probability value** (or ***P*-value**) that is associated with the test statistic.

THE STATISTICAL NULL HYPOTHESIS Before we can define the probability of a statistical test, we must first define the statistical null hypothesis, or H_0. We noted above that scientists favor parsimonious or simple explanations over more complex ones. What is the simplest explanation to account for the difference in the means of the two groups? In our example, the simplest explanation is that the differences represent random variation between the groups and do not reflect any systematic effect of snowmobiles. In other words, if we were to divide the wolves into two groups but not expose individuals in either

[12] Many people try to answer this question by simply comparing the means. However, we cannot evaluate a difference between means unless we also have some feeling for how much individuals within a treatment group differ. For example, if several of the individuals in the no-snowmobile group have GC levels as low as 200 ng/g and others have GC levels as high as 1544 ng/g (the average, remember, was 872), then a difference of 600 ng/g between the two exposure groups may not mean much. On the other hand, if most individuals in the no-snowmobile group have GC levels between 850 and 950 ng/g, then a 600 ng/g difference is substantial. As we discussed in Chapter 3, we need to know not only the difference in the means, but the variance about those means—the amount that a typical individual differs from its group mean. Without knowing something about the variance, we cannot say anything about whether differences between the means of two groups are meaningful.

group to snowmobiles, we might still find that the means differ from each other. Remember that it is extremely unlikely that the means of two samples of numbers will be the same, even if they were sampled from a larger population using an identical process.

Glucocorticoid levels will differ from one individual to another for many reasons that cannot be studied or controlled in this experiment, and all of this variation—including variation due to measurement error—is what we label random variation. We want to know if there is any evidence that the observed difference in the mean GC levels of the two groups is larger than we would expect given the random variation among individuals. Thus, a typical statistical null hypothesis is that "differences between groups are no greater than we would expect due to random variation." We call this a **statistical null hypothesis** because the hypothesis is that a specific mechanism or force—some force *other* than random variation—does *not* operate.

THE ALTERNATIVE HYPOTHESIS Once we state the statistical null hypothesis, we then define one or more *alternatives* to the null hypothesis. In our example, the natural **alternative hypothesis** is that the observed difference in the average GC levels of the two groups is too large to be accounted for by random variation among individuals. Notice that the alternative hypothesis is *not* that snowmobile exposure is responsible for an increase in GC! Instead, the alternative hypothesis is focused simply on the *pattern* that is present in the data. The investigator can *infer* mechanism from the pattern, but that inference is a separate step. The statistical test merely reveals whether the pattern is likely or unlikely, given that the null hypothesis is true. Our ability to assign causal mechanisms to those statistical patterns depends on the quality of our experimental design and our measurements.

For example, suppose the group of wolves exposed to snowmobiles had also been hunted and chased by humans and their hunting dogs within the last day, whereas the unexposed group included wolves from a remote area uninhabited by humans. The statistical analysis would probably reveal significant differences in GC levels between the two groups regardless of exposure to snowmobiles. However, it would be dangerous to conclude that the difference between the means of the two groups was caused by snowmobiles, even though we can reject the statistical null hypothesis that the pattern is accounted for by random variation among individuals. In this case, the treatment effect is **confounded** with other differences between the control and treatment groups (exposure to hunting dogs) that are potentially related to stress levels. As we will discuss in

Chapters 6 and 7, an important goal of good experimental design is to avoid such confounded designs.

If our experiment was designed and executed correctly, it may be safe to infer that the difference between the means is caused by the presence of snowmobiles. But even here, we cannot pin down the precise physiological mechanism if all we did was measure the GC levels of exposed and unexposed individuals. We would need much more detailed information on hormone physiology, blood chemistry, and the like if we want to get at the underlying mechanisms.[13] Statistics help us establish convincing patterns, and from those patterns we can begin to draw inferences or conclusions about cause-and-effect relationships.

In most tests, the alternative hypothesis is not explicitly stated because there is usually more than one alternative hypothesis that could account for the patterns in the data. Rather, we consider the set of alternatives to be "not H_0." In a Venn diagram, all outcomes of data can then be classified into either H_0 or not H_0.

THE *P*-VALUE In many statistical analyses, we ask whether the null hypothesis of random variation among individuals can be rejected. The *P*-value is a guide to making that decision. A statistical *P*-value measures the probability that observed or more extreme differences would be found *if the null hypothesis were true.* Using the notation of conditional probability introduced in Chapter 1, $P\text{-value} = P(\text{data} \mid H_0)$.

Suppose the *P*-value is relatively small (close to 0.0). Then it is unlikely (the probability is small) that the observed differences could have been obtained if the null hypothesis were true. In our example of wolves and snowmobiles, a low *P*-value would mean that it is unlikely that a difference of 600 ng/g in GC levels would have been observed between the exposed and unexposed groups if there was only random variation among individuals and no consistent effect of snowmobiles (i.e., if the null hypothesis is true). Therefore, with a small *P*-value, the results would be improbable given the null hypothesis, so we reject it. Because we had only one alternative hypothesis in our study, our conclusion is that snow-

[13] Even if the physiological mechanisms were elucidated, there would still be questions about ultimate mechanisms at the molecular or genetic level. Whenever we propose a mechanism, there will always be lower-level processes that are not completely described by our explanation and have to be treated as a "black box." However, not all higher-level processes can be explained successfully by reductionism to lower-level mechanisms.

mobiles (or something associated with them) could be responsible for the difference between the treatment groups.[14]

On the other hand, suppose that the calculated *P*-value is relatively large (close to 1.0). Then it is likely that the observed difference could have occurred given

[14] Accepting an alternative hypothesis based on this mechanism of testing a null hypothesis is an example of the fallacy of "affirming the consequent" (Barker 1989). Formally, the *P*-value = $P(\text{data} \mid H_0)$. If the null hypothesis is true, it would result in (or in the terms of logic, *imply*) a particular set of observations (here, the data). We can write this formally as $H_0 \Rightarrow$ null data, where the arrow is read as "implies." If your observations are different from those expected under H_0, then a low *P*-value suggests that $H_0 \not\Rightarrow$ your data, where the crossed arrow is read as "does not imply." Because you have set up only one alternative hypothesis, H_a, then you are further asserting that $H_a = \neg H_0$ (where the symbol $\neg$ means "not"), and the only possibilities for data are those data possible under H_0 ("null data") and those not possible under H_0 ("$\neg$null data" = "your data"). Thus, you *are asserting the following logical progression:*

1. Given: $H_0 \Rightarrow$ null data
2. Observe: $\neg$null data
3. Conclude: $\neg$null data $\Rightarrow \neg H_0$
4. Thus: $\neg H_0 (= H_a) \Rightarrow \neg$null data

But really, all you can conclude is point 3: $\neg$null data $\Rightarrow \neg H_0$ (the so-called *contrapositive* of 1). In 3, the alternative hypothesis (H_a) is the "consequent," and you cannot assert its truth simply by observing its "predicate" ($\neg$null data in 3); many other possible causes could have yielded your results ($\neg$null data). You can affirm the consequent (assert H_a is *true*) *if and only if there is only one possible alternative to your null hypothesis.* In the simplest case, where H_0 asserts "no effect" and H_a asserts "some effect," proceeding from 3 to 4 makes sense. But biologically, it is usually of more interest to know what is the actual effect (as opposed to simply showing there is "some effect").

Consider the ant example earlier in this chapter. Let H_0 = all 25 ants in the Harvard Forest are *Myrmica*, and H_a = 10 ants in the forest are not *Myrmica*. If you collect a specimen of *Camponotus* in the forest, you can conclude that the data imply that the null hypothesis is false (observation of a *Camponotus* in the forest $\Rightarrow \neg H_0$). But you cannot draw any conclusion about the alternative hypothesis. You could support a less stringent alternative hypothesis, H_a = not all ants in the forest are *Myrmica*, but affirming this alternative hypothesis does not tell you anything about the actual distribution of ants in the forest, or the identity of the species and genera that are present.

This is more than splitting logical hairs. Many scientists appear to believe that when they report a *P*-value that they are giving the probability of observing the null hypothesis given the data [$P(H_0 \mid \text{data})$] or the probability that the alternative hypothesis is false, given the data [$1 - P(H_a \mid \text{data})$]. But, in fact, they are reporting something completely different—the probability of observing the data given the null hypothesis: $P(\text{data} \mid H_0)$. Unfortunately, as we saw in Chapter 1, $P(\text{data} \mid H_0) \neq P(H_0 \mid \text{data}) \neq 1 - P(H_a \mid \text{data})$; in the words of the immortal anonymous philosopher from Maine, *you can't get there from here.* However, it is possible to compute directly $P(H_0 \mid \text{data})$ or $P(H_a \mid \text{data})$ using Bayes' Theorem (Chapter 1) and the Bayesian methods outlined in Chapter 5.

that the null hypothesis is true. In this example, a large *P*-value would mean that a 600-ng/g difference in GC levels likely would have been observed between the exposed and unexposed groups even if snowmobiles had no effect and there was only random variation among individuals. That is, with a large *P*-value, the observed results would be likely under the null hypothesis, so we do not have sufficient evidence to reject it. Our conclusion is that differences in GC levels between the two groups can be most parsimoniously attributed to random variation among individuals.

Keep in mind that when we calculate a statistical *P*-value, we are viewing the data through the lens of the null hypothesis. If the patterns in our data are likely under the null hypothesis (large *P*-value), we have no reason to reject the null hypothesis in favor of more complex explanations. On the other hand, if the patterns are unlikely under the null hypothesis (small *P*-value), it is more parsimonious to reject the null hypothesis and conclude that something more than random variation among subjects contributes to the results.

WHAT DETERMINES THE *P*-VALUE? The calculated *P*-value depends on three things: the number of observations in the samples (n), the difference between the means of the samples ($\bar{Y}_i - \bar{Y}_j$), and the level of variation among individuals (s^2). The more observations in a sample, the lower the *P*-value, because the more data we have, the more likely it is we are estimating the true population means and can detect a real difference between them, if it exists (see the Law of Large Numbers in Chapter 3). The *P*-value also will be lower the more different the two groups are in the variable we are measuring. Thus, a 10-ng/g difference in mean GC levels between control and treatment groups will generate a lower *P*-value than a 2-ng/g difference, all other things being equal. Finally, the *P*-value will be lower if the variance among individuals within a treatment group is small. The less variation there is from one individual to the next, the easier it will be to detect differences among groups. In the extreme case, if the GC levels for all individuals within the group of wolves exposed to snowmobiles were identical, and the GC levels for all individuals within the unexposed group were identical, then any difference in the means of the two groups, no matter how small, would generate a low *P*-value.

WHEN IS A *P*-VALUE SMALL ENOUGH? In our example, we obtained a *P*-value = 0.01 for the probability of obtaining the observed difference in GC levels between wolves exposed to and not exposed to snowmobiles. Thus, if the null hypothesis were true and there was only random variation among individuals in the data, the chance of finding a 600-ng/g difference in GC between exposed and unexposed groups is only 1 in 100. Stated another way, if the null hypothesis were

true, and we conducted this experiment 100 times, using different subjects each time, in only *one* of the experiments would we expect to see a difference as large or larger than what we actually observed. Therefore, it seems unlikely the null hypothesis is true, and we reject it. If our experiment was properly designed, we can safely conclude that snowmobiles cause increases in GC levels, although we cannot specify what it is about snowmobiles that causes this response. On the other hand, if the calculated statistical probability were $P = 0.88$, then we would expect a result similar to what we found in 88 out of 100 experiments due to random variation among individuals; our observed result would not be at all unusual under the null hypothesis, and there would be no reason to reject it.

But what is the precise cutoff point that we should use in making the decision to reject or not reject the null hypothesis? This is a judgment call, as there is no natural critical value below which we should always reject the null hypothesis and above which we should never reject it. However, after many decades of custom, tradition, and vigilant enforcement by editors and journal reviewers, the operational critical value for making these decisions equals 0.05. In other words, if the statistical probability $P \leq 0.05$, the convention is to reject the null hypothesis, and if the statistical probability $P > 0.05$, the null hypothesis is not rejected. When scientists report that a particular result is "significant," they mean that they rejected the null hypothesis with a P-value ≤ 0.05.[15]

A little reflection should convince you that a critical value of 0.05 is relatively low. If you used this rule in your everyday life, you would never take an umbrella with you unless the forecast for rain was at least 95%. You would get wet a lot more often than your friends and neighbors. On the other hand, if your friends and neighbors saw you carrying your umbrella, they could be pretty confident of rain.

In other words, setting a critical value = 0.05 as the standard for rejecting a null hypothesis is very conservative. We require the evidence to be very strong in order to reject the statistical null hypothesis. Some investigators are unhappy about using an arbitrary critical value, and about setting it as low as 0.05. After all, most of us would take an umbrella with a 90% forecast of rain, so why shouldn't we be a bit less rigid in our standard for rejecting the null hypothesis? Perhaps we should set the critical critical value = 0.10, or perhaps we should use different critical values for different kinds of data and questions.

[15] When scientists discuss "significant" results in their work, they are really speaking about how confident they are that a statistical null hypothesis has been correctly rejected. But the public equates "significant" with "important." This distinction causes no end of confusion, and it is one of the reasons that scientists have such a hard time communicating their ideas clearly in the popular press.

A defense of the 0.05-cutoff is the observation that scientific standards need to be high so that investigators can build confidently on the work of others. If the null hypothesis is rejected with more liberal standards, there is a greater risk of falsely rejecting a true null hypothesis (a Type I error, described in more detail below). If we are trying to build hypotheses and scientific theories based on the data and results of others, such mistakes slow down scientific progress. By using a low critical value, we can be confident that the patterns in the data are quite strong. However, even a low critical value is not a safeguard against a poorly designed experiment or study. In such cases, the null hypothesis may be rejected, but the patterns in the data reflect flaws in the sampling or manipulations, not underlying biological differences that we are seeking to understand.

Perhaps the strongest argument in favor of requiring a low critical value is that we humans are psychologically predisposed to recognizing and seeing patterns in our data, even when they don't exist. Our vertebrate sensory system is adapted for organizing data and observations into "useful" patterns, generating a built-in bias towards rejecting null hypotheses and seeing patterns where there is really randomness (Sale 1984).[16] A low critical value is a safeguard against such activity. A low critical value also helps act as a gatekeeper on the rate of scientific publications because non-significant results are much less likely to be reported or published.[17] We emphasize, however, that no law *requires* a critical value to be ≤ 0.05 in order for the results to be declared significant. In many cases, it may be more useful to report the exact P-value and let the readers decide for themselves how important the results are. However, the practical reality is that reviewers and editors will usually not allow you to discuss mechanisms that are not supported by a $P \leq 0.05$ result.

[16] A fascinating illustration of this is to ask a friend to draw a set of 25 randomly located points on a piece of paper. If you compare the distribution of those points to a set of truly random points generated by a computer, you will often find that the drawings are distinctly non-random. People have a tendency to space the points too evenly across the paper, whereas a truly random pattern generates apparent "clumps" and "holes." Given this tendency to see patterns everywhere, we should use a low critical value to ensure we are not deceiving ourselves.

[17] The well-known tendency for journals to reject papers with non-significant results (Murtaugh 2002a) and authors to therefore not bother trying to publish them is not a good thing. In the hypothetico-deductive method, science progresses through the elimination of alternative hypotheses, and this can often be done when we fail to reject a null hypothesis. However, this approach requires authors to specify and test the unique predictions that are made by competing alternative hypotheses. Statistical tests based on H_0 versus not H_0 do not often allow for this kind of specificity.

STATISTICAL HYPOTHESES VERSUS SCIENTIFIC HYPOTHESES REDUX The biggest difficulty in using *P*-values results from the failure to distinguish statistical null hypotheses from scientific hypotheses. Remember that a *scientific hypothesis* poses a formal mechanism to account for patterns in the data. In this case, our scientific hypothesis is that snowmobiles cause stress in wolves, which we propose to test by measuring GC levels. Higher levels GC might come about by complex changes in physiology that lead to changes in GC production when an animal is under stress. In contrast, the *statistical null hypothesis* is a statement about patterns in the data and the likelihood that these patterns could arise by chance or random processes that are not related to the factors we are explicitly studying.

We use the methods of probability when deciding whether or not to reject the statistical null hypothesis; think of this process as a method for establishing pattern in the data. Next, we draw a conclusion about the validity of our scientific hypothesis based on the statistical pattern in this data. The strength of this inference depends very much on the details of the experiment and sampling design. In a well-designed and replicated experiment that includes appropriate controls and in which individuals have been assigned randomly to clear-cut treatments, we can be fairly confident about our inferences and our ability to evaluate the scientific hypothesis we are considering. However, in a sampling study in which we have not manipulated any variables but have simply measured differences among groups, it is difficult to make solid inferences about the underlying scientific hypotheses, even if we have rejected the statistical null hypothesis.[18]

We think the more general issue is not the particular critical value that is chosen, but whether we always should be using an hypothesis-testing framework. Certainly, for many questions statistical hypothesis tests are a powerful way to establish what patterns do or do not exist in the data. But in many studies, the real issue may not be hypothesis testing, but **parameter estimation.** For example, in the stress study, it may be more important to determine the range of GC levels expected for wolves exposed to snowmobiles rather than merely to establish that snowmobiles significantly increases GC levels. We also should establish the level of confidence or certainty in our parameter estimates.

[18] In contrast to the example of the snowmobiles and wolves, suppose we measured the GC levels of 10 randomly chosen old wolves and 10 randomly chosen young ones. Could we be as confident about our inferences as in the snowmobile experiment? Why or why not? What are the differences, if any, between experiments in which we manipulate individuals in different groups (exposed wolves versus unexposed wolves) and sampling surveys in which we measure variation among groups but do not directly manipulate or change conditions for those groups (old wolves versus young wolves)?

Errors in Hypothesis Testing

Although statistics involves many precise calculations, it is important not to lose sight of the fact that statistics is a discipline steeped in uncertainty. We are trying to use limited and incomplete data to make inferences about underlying mechanisms that we may understand only partially. In reality, the statistical null hypothesis is either true or false; if we had complete and perfect information, we would know whether or not it were true and we would not need statistics to tell us. Instead, we have only our data and methods of statistical inference to decide whether or not to reject the statistical null hypothesis. This leads to an interesting 2×2 table of possible outcomes whenever we test a statistical null hypothesis (Table 4.1)

Ideally, we would like to end up in either the upper left or lower right cells of Table 4.1. In other words, when there is only random variation in our data, we would hope to not reject the statistical null hypothesis (upper left cell), and when there is something more, we would hope to reject it (lower right cell). However, we may find ourselves in one of the other two cells, which correspond to the two kinds of errors that can be made in a statistical decision.

TYPE I ERROR If we falsely reject a null hypothesis that is true (upper right cell in Table 4.1), we have made a false claim that some factor above and beyond random variation is causing patterns in our data. This is a Type I error, and by convention, the probability of committing a Type I error is denoted by α. When you calculate a statistical *P*-value, you are actually estimating α. So, a more precise

TABLE 4.1 The quadripartite world of statistical testing

	Retain H_0	Reject H_0
H_0 true	Correct decision	Type I error (α)
H_0 false	Type II error (β)	Correct decision

Underlying null hypotheses are either true or false, but in the real world we must use sampling and limited data to make a decision to accept or reject the null hypothesis. Whenever a statistical decision is made, one of four outcomes will result. A correct decision results when we retain a null hypothesis that is true (upper left-hand corner) or reject a null hypothesis that is false (lower right-hand corner). The other two possibilities represent errors in the decision process. If we reject a null hypothesis that is true, we have committed a Type I error (upper right-hand corner). Standard parametric tests seek to control α, the probability of a Type I error. If we retain a null hypothesis that is false, we have committed a Type II error (lower left-hand corner). The probability of a Type II error is β.

definition of a *P*-value is that it is the chance we will make a Type I error by falsely rejecting a true null hypothesis.[19] This definition lends further support for asserting statistical significance only when *P*-value is very small. The smaller the *P*-value, the more confident we can be that we will not commit a Type I error if we reject H_0. In the glucocorticoid example, the risk of making a Type I error by rejecting the null hypothesis is 1%. As we noted before, scientific publications use a standard of a maximum of a 5% risk of Type I error for rejecting a null hypothesis. In environmental impact assessment, a Type I error would be a "false positive" in which, for example, an effect of a pollutant on human health is reported but does not, in fact, exist.

TYPE II ERROR AND STATISTICAL POWER The lower left cell in Table 4.1 represents a Type II error. In this case, the investigator has incorrectly failed to reject a null hypothesis that is false. In other words, there are systematic differences between the groups being compared, but the investigator has failed to reject the null hypothesis and has concluded incorrectly that only random variation among observations is present. By convention, the probability of committing a Type II error is denoted by β. In environmental assessment, a Type II error would be a "false negative" in which, for example, there is an effect of a pollutant on human health, but it is not detected.[20]

[19] We have followed standard statistical treatments that equate the calculated *P*-value with the estimate of Type I error rate α. However, Fisher's evidential *P*-value may not be strictly equivalent to Neyman and Pearson's α. Statisticians disagree whether the distinction is an important philosophical issue or simply a semantic difference. Hubbard and Bayarri (2003) argue that the incompatibility is important, and their paper is followed by discussion, comments, and rebuttals from other statisticians. Stay tuned!

[20] The relationship between Type I and Type II errors informs discussions of the precautionary principle of environmental decision making. Historically, for example, regulatory agencies have assumed that new chemical products were benign until proven harmful. Very strong evidence was required to reject the null hypothesis of no effect on health and well-being. Manufacturers of chemicals and other potential pollutants are keen to minimize the probability of committing Type I error. In contrast, environmental groups that serve the general public are interested in minimizing the probability that the manufacturer committed a Type II error. Such groups assume that a chemical is harmful until proven benign, and are willing to accept a larger probability of committing a Type I error if this means they can be more confident that the manufacturer has not falsely accepted the null hypothesis. Following such reasoning, in assessing quality control of industrial production, Type I and Type II errors are often known as producer and consumer errors, respectively (Sokal and Rohlf 1995).

A concept related to the probability of committing a Type II error is the **power** of a statistical test. Power is calculated as $1 - \beta$, and equals the probability of correctly rejecting the null hypothesis when it is false. We want our statistical tests to have good power so that we have a good chance of detecting significant patterns in our data when they are present.

WHAT IS THE RELATIONSHIP BETWEEN TYPE I AND TYPE II ERROR? Ideally, we would like to minimize both Type I and Type II errors in our statistical inference. However, strategies designed to reduce Type I error inevitably increase the risk of Type II error, and vice versa. For example, suppose you decide to reject the null hypothesis only if $P < 0.01$—a fivefold more stringent standard than the conventional criterion of $P < 0.05$. Although your risk of committing a Type I error is now much lower, there is a much greater chance that when you fail to reject the null hypothesis, you may be doing so incorrectly (i.e., you will be committing a Type II error). Although Type I and Type II errors are inversely related to one another, there is no simple mathematical relationship between them, because the probability of a Type II error depends in part on what the alternative hypothesis is, how large an effect we hope to detect (Figure 4.5), the sample size, and the wisdom of our experimental design or sampling protocol.

WHY ARE STATISTICAL DECISIONS BASED ON TYPE I ERROR? In contrast to the probability of committing a Type I error, which we determine with standard statistical tests, the probability of committing a Type II error is not often calculated or reported, and in many scientific papers, the probability of committing a Type II error is not even discussed. Why not? To begin with, we often cannot calculate the probability of a Type II error unless the alternative hypotheses are completely specified. In other words, if we want to determine the risk of falsely accepting the null hypothesis, the alternatives have to be fleshed out more than just "not H_0." In contrast, calculating the probability of a Type I error does not require this specification; instead we are required only to meet some assumptions of normality and independence (see Chapters 9 and 10).

On a philosophical basis, some authors have argued that a Type I error is a more serious mistake in science than a Type II error (Shrader-Frechette and McCoy 1992). A Type I error is an error of falsity, in which we have incorrectly rejected a null hypothesis and made a claim about a more complex mechanism. Others may follow our work and try to build their own studies based on that false claim. In contrast, a Type II error is an error of ignorance. Although we have not rejected the null hypothesis, someone else with a better experiment or more data may be able to do so in the future, and the science will progress

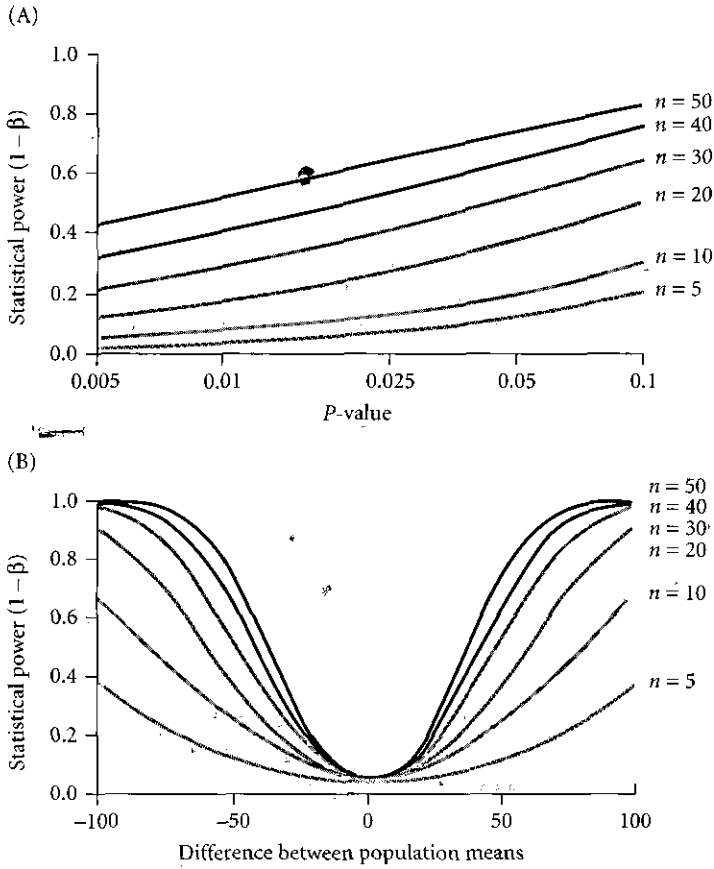
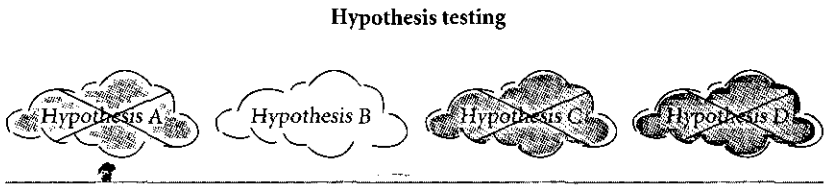


Figure 4.5 The relationship between statistical power, *P*-values, and observable effect sizes as a function of sample size. (A) The *P*-value is the probability of incorrectly rejecting a true null hypothesis, whereas statistical power is the probability of correctly rejecting a false null hypothesis. The general result is that the lower the *P*-value used for rejection of the null hypothesis, the lower the statistical power of correctly detecting a treatment effect. At a given *P*-value, statistical power is greater when the sample size is larger. (B) The smaller the observable effect of the treatment (i.e., the smaller the difference between the treatment group and the control group), the larger the sample size necessary for good statistical power to detect a treatment effect.[21]

from that point. However, in many applied problems, such as environmental monitoring or disease diagnosis, Type II errors may have more serious consequences because diseases or adverse environmental effects would not be correctly detected.

[21] We can apply these graphs to our example comparing glucocorticoid hormone levels for populations of wolves exposed to snowmobiles (treatment group) versus wolves that were not exposed to snowmobiles (control group). In the original data (Creel et al. 2002), the standard deviation of the control population of wolves unexposed to snowmobiles was 73.1, and that of wolves exposed to snowmobiles was 114.2. Panel (A) suggests that if there were a 50-ng/g difference between the experimental populations, and if sample size was $n = 50$ in each group, the experimenters would have correctly accepted the alternative hypothesis only 51% of the time for $P = 0.01$. Panel (B) shows that power increases steeply as the populations become more different. In the actual well-designed study (Creel et al. 2002), sample size was 193 in the unexposed group and 178 in the exposed group, the difference between population means was 598 ng/g, and the actual power of the statistical test was close to 1.0 for $P = 0.01$.



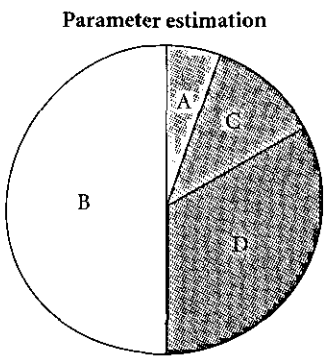


Figure 4.6 Hypothesis testing versus parameter estimation. Parameter estimation more easily accommodates multiple mechanisms and may allow for an estimate of the relative importance of different factors. Parameter estimation may involve the construction of confidence or credibility intervals (see Chapter 3) to estimate the strength of an effect. A related technique in the analysis of variance is to decompose the total variation in the data into proportions that are explained by different factors in the model (see Chapter 10). Both methods quantify the relative importance of different factors, whereas hypothesis testing emphasizes a binary yes/no decision as to whether a factor has a measurable effect or not.

Parameter Estimation and Prediction

All the methods for hypothesis testing that we have described—the inductive method (and its modern descendant, Bayesian inference), the hypothetico-deductive method, and statistical hypothesis testing are concerned with choosing a single explanatory "answer" from an initial set of multiple hypotheses. In ecology and environmental science, it is more likely that many mechanisms may be operating simultaneously to produce observed patterns; an hypothesis-testing framework that emphasizes single explanations may not be appropriate. Rather than try to test multiple hypotheses, it may be more worthwhile to estimate the relative contributions of each to a particular pattern. This approach is sketched in Figure 4.6, in which we partition the effects of each hypothesis on the observed patterns by estimating how much each cause contributes to the observed effect.

In such cases, rather than ask whether a particular cause has some effect versus no effect (i.e., is it significantly different from 0.0?), we ask what is the best estimate of the **parameter** that expresses the magnitude of the effect.[22] For example, in Figure 4.3, measured photosynthetic rates for young sun leaves of

[22] Chapter 9 will introduce some of the strategies used for fitting curves and estimating parameters from data. See Hilborn and Mangel (1997) for a detailed discussion. Clark et al. (2003) describe recent Bayesian strategies to curve fitting.

Rhizophora mangle were fit to a Michaelis-Menten equation. This equation is a simple model that describes a variable rising smoothly to an asymptote. The Michaelis-Menten equation shows up frequently in biology, being used to describe everything from enzyme kinematics (Real 1977) to invertebrate foraging rates (Holling 1959).

The Michaelis-Menten equation takes the form

$$Y = \frac{kX}{X + D}$$

where k and D are the two fitted parameters of the model, and X and Y are the independent and dependent variables. In this example, k represents the asymptote of the curve, which in this case is the maximum assimilation rate; the independent variable X is light intensity; and the dependent variable Y is the net assimilation rate. For the data in Figure 4.3, the parameter estimate for k is a maximum assimilation rate of 7.1 μmol CO_2 m^{-2} s^{-1}. This accords well with an "eyeball estimate" of where the asymptote would be on this graph.

The second parameter in the Michaelis-Menten equation is D, the half-saturation constant. This parameter gives the value of the X variable that yields a Y variable that is half of the asymptote. The smaller D is, the more quickly the curve rises to the asymptote. For the data in Figure 4.3, the parameter estimate for D is photosynthetically active radiation (PAR) of 250 μmol CO_2 m^{-2} s^{-1}.

We also can measure the uncertainty in these parameter estimates by using estimates of standard error to construct confidence or credibility intervals (see Chapter 3). The estimated standard error for $k = 0.49$, and for $D = 71.3$. Statistical hypothesis testing and parameter estimation are related, because if the confidence interval of uncertainty includes 0.0, we usually are not able to reject the null hypothesis of no effect for one of the mechanisms. For the parameters k and D in Figure 4.3, the P-values for the test of the null hypothesis that the parameter does not differ from 0.0 are 0.0001 and 0.004, respectively. Thus, we can be fairly confident in our statement that these parameters are greater than 0.0. But, for the purposes of evaluating and fitting models, the numerical values of the parameters are more informative than just asking whether they differ or not from 0.0. In later chapters, we will give other examples of studies in which model parameters are estimated from data.

Summary

Science is done using inductive and hypothetico-deductive methods. In both methods, observed data are compared with data predicted by the hypotheses. Through the inductive method, which includes modern Bayesian analyses, a

single hypothesis is repeatedly tested and modified; the goal is to confirm or assert the probability of a particular hypothesis. In contrast, the hypothetico-deductive method requires the simultaneous statement of multiple hypotheses. These are tested against observations with the goal of falsifying or eliminating all but one of the alternative hypotheses. Statistics are used to test hypotheses objectively, and can be used in both inductive and hypothetico-deductive approaches.

Probabilities are calculated and reported with virtually all statistical tests. The probability values associated with statistical tests may allow us to infer causes of the phenomena that we are studying. Tests of statistical hypotheses using the hypothetico-deductive method yield estimates of the chance of obtaining a result equal to or more extreme than the one observed, given that the null hypothesis is true. This *P*-value is also the probability of incorrectly rejecting a true null hypothesis (or committing a Type I statistical error). By convention and tradition, 0.05 is the cutoff value in the sciences for claiming that a result is statistically significant. The calculated *P*-value depends on the number of observations, the difference between the means of the groups being compared, and the amount of variation among individuals within each group. Type II statistical errors occur when a false null hypothesis is incorrectly accepted. This kind of error may be just as serious as a Type I error, but the probability of Type II errors is reported rarely in scientific publications. Tests of statistical hypotheses using inductive or Bayesian methods yield estimates of the probability of the hypothesis or hypotheses of interest given the observed data. Because these are confirmatory methods, they do not give probabilities of Type I or Type II errors. Rather, the results are expressed as the odds or likelihood that a particular hypothesis is correct.

Regardless of the method used, all science proceeds by articulating testable hypotheses, collecting data that can be used to test the predictions of the hypotheses, and relating the results to underlying cause-and-effect relationships.