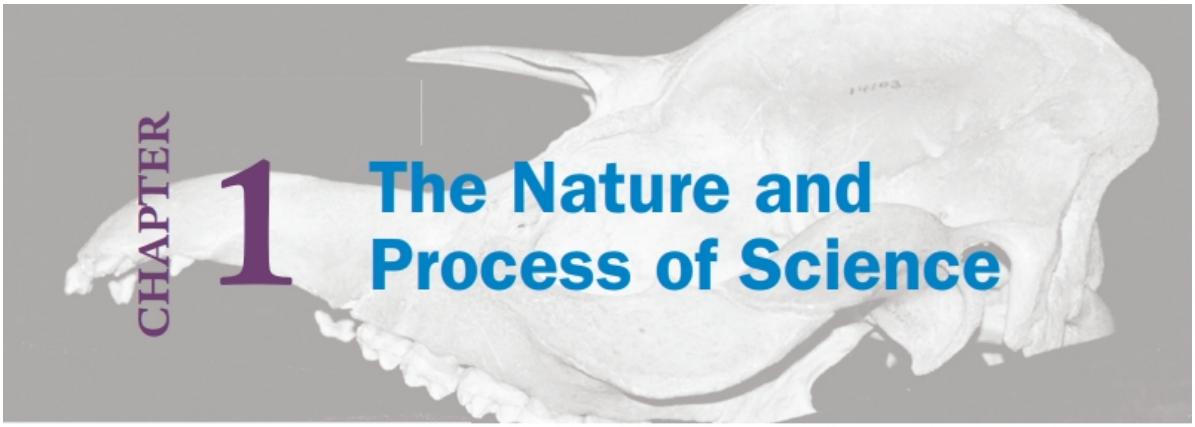

CHAPTER 1

The Nature and Process of Science

This chapter outlines the scientific method in a way that sets up how we can apply this method to evolutionary concepts. Our understanding of the natural world consists of the limited number of things we observe and our conjectures as to what else fills in the gaps. New scientific observations are being made all the time, many of them as a result of laboratory experiments and field studies. The conjectures are our hypotheses of the causes behind the patterns we perceive in the natural world, as well as our predictions of what else we will observe when we make further investigations. Good scientific hypotheses make predictions about what we should find in the world, and we can test hypotheses by seeing if their predictions match what we observe, whether these observations are the results of a laboratory experiment or discoveries in the fossil record.

1.1 Questions, patterns, and the natural world

We usually begin discussion of the scientific method with making an **observation** that stimulates the formation of a **question**. This is actually one of the biggest stumbling blocks for students, as they feel unable to identify questions to investigate. In fact, our daily lives are filled with questions to which we can apply scientific thinking. Issues as mundane as a car that won't start lead you to wonder why, then to hypothesize what might be wrong, and subsequently to think of other observations you can make to test that hypothesis.

Questions can come from things that directly affect our lives, but most scientists ask questions about broader patterns that we perceive in the natural world. Why do plants grow? Why are there seasons? As we think more about natural phenomena, new questions present themselves. In many cases, more specific questions may lead a scientist to ask more general questions. Why do polar bears live in the Arctic but not in Antarctica? This might require answering the broader question of why the poles don't have the same species living in them when they are such similar environments.

Brainstorm

Working in groups, come up with as many questions as you can regarding natural phenomena. (Remember: natural phenomena include everything that we can observe, so even things you might not think of as "natural"—things made by or concerning humans, for instance—are actually part of the natural world.)

1.2 Hypotheses

Questions naturally lead to wondering about answers, and our attempted answers are **hypotheses**. Good hypotheses have certain qualities. They are usually about **causation**; in other words, they propose causes or reasons for why a phenomenon exists. Returning to the example of why a car won't start, a possible reason that this is true—the car is out of gas or the battery is dead—would be a hypothesis.

When we think of causes, we might be interested in whether two things that we can measure are related (or **correlated**, as we'll discuss in a later chapter), where, say, when X is high, Y is also high, and when X is low, Y is also low. But is this relationship causal? Does X determine Y, or vice versa (or neither)? For instance, we might find our car that won't start also has a radio that doesn't function. That doesn't necessarily mean that the radio is the reason the car won't start. Good hypotheses typically try to get at these causes, rather than just vague statements of relationship.

Good hypotheses are **testable** and **falsifiable**; they "stick their necks out," so to speak, and can be subjected to testing that can result in rejection. Why do we value the possibility of rejection over the possibility of proof? This is because it is logically possible to disprove a hypothesis but not to prove it, a concept that comes from the work of the philosopher of science, Karl Popper. For instance, take the hypothesis "X causes Y." I could then run an experiment where I measure X and Y and predict, based on the hypothesis, that I should get high values for Y when X is high, and low values for Y when X is low. If I get no such relationship between X and Y, I can reject my hypothesis. If I get the predicted relationship, it is consistent with my hypothesis, but have I truly ruled out the possibility that some third phenomenon (call it Z) is driving the values of X and Y?

Again, we can illustrate this with our example of the car that won't start. If my hypothesis is that the battery is dead, I can make predictions about other things that should be true if the battery is dead, such as that the radio should not work. If the radio doesn't work, that supports my hypothesis, but it doesn't rule out the possibility that the radio is not working for some other reason. On the other hand, if the radio does work, then I positively know that the battery can't be dead. A working radio falsifies the dead battery hypothesis.

We can falsify hypotheses, but if we can't actually prove a hypothesis, what, then, allows us to elevate a hypothesis to being an accepted explanation for something? It has to consistently defy rejection, and it has to do so better than competing hypotheses.

Ultimately, we want to compare hypotheses to see which one is better supported by our data. Even if we haven't thought of two separate possible answers to our question, we can still compare our one hypothesis to the **null hypothesis**. The null hypothesis is essentially the negation of the hypothesis we are considering (which we will call the **alternate hypothesis**). That does not mean that the null hypothesis is the opposite of the alternate hypothesis; rather, if we are hypothesizing that X causes Y, our null hypothesis would be that X does not cause Y, not that Y causes X. For instance, if we were considering the alternate hypothesis that hot dogs cause earthquakes, our null hypothesis would be that hot dogs do not cause earthquakes, or that hot dogs have nothing to do with earthquakes. It would not be that earthquakes cause hot dogs!

Since we will be referring to hypotheses very often, we will use the letter H for "hypothesis," with a subscript to identify a specific hypothesis. Often, we'll simply use H_A to represent the alternate hypothesis and H_0 (subscript "zero") for the null hypothesis.

Activity

Come up with hypotheses to answer some of the questions you developed in the brainstorming in the previous section, or consider sample hypotheses provided by your instructor. Are they testable? Are they falsifiable?

Come up with null hypotheses for your hypotheses, or for hypotheses provided by your instructor.

1.3 Predictions

Presumably our hypothesis does a good job of explaining whatever we observed that caused us to ask a question; in other words, if it were true, our hypothesis would fully answer our question. We can make **predictions** based on that hypothesis regarding what else we might observe. Ideally, if those predictions don't match what we observe, we can reject our hypothesis. In the car example, we made a prediction about the functioning of the radio based on the hypothesis that a dead battery was preventing the car from starting; we also could have made a prediction that the radio would work based on the null hypothesis that the battery had nothing to do with why the car wouldn't start. In reality, data rarely fit a hypothesis perfectly, due to various sources of error, so what we do instead is compare how competing hypotheses fit our data and reject the hypothesis that has the worst fit. If we're only considering a single hypothesis, we can compare its predictions to those of the null hypothesis. When we do an experiment, we essentially attempt to manipulate conditions such that results predicted by the hypotheses being compared will be different.

For instance, consider the hypothesis that plants need sunlight to grow. The null hypothesis would be that sunlight has no effect on plant growth. We could then make predictions about how plants will grow with and without sunlight. According to our alternate hypothesis, plants would grow much better in sunlight than they would in darkness. Our null hypothesis would predict that plants will grow about the same regardless of whether they have sunlight or not. Thus, the alternate and null hypotheses predict that we will observe different things depending on which hypothesis is true.

Note that when we make predictions, we often need to specify exactly what the observations would be and how we would measure them. For instance, in our example above, we made predictions about growth based on each hypothesis. But what do we mean by "growth"? How would we know if one plant has "grown" more than another, and how could we represent this to another scientist? To solve this, we could take a measurement that we think describes growth. We could measure the height of a plant, or count its leaves, or weigh it. But "growth" refers to how much the plant has changed during the experiment,

so we could take the measurement at the beginning and the end of the experiment and determine the difference between the two; the value of the difference would be our measurement of growth. What we have constructed is a measurement to use as our **response variable**; in other words, this is the variable that we expect to have different values (i.e., to "respond" differently) in the experiment depending on whether the alternate or null hypothesis is true.

The response variable can also be called the **dependent variable**, because, at least according to the alternate hypothesis, its value should be dependent on another variable. In the case of the plant experiment, our dependent variable of growth is determined by the extent to which the plant is exposed to sunlight. The amount of sunlight would be our **independent variable**. You will often see graphs depicting the relationship between an independent and a dependent variable; the convention is for the *x* axis to represent the independent variable and the *y* axis to represent the dependent variable.

Activity

For each of the examples of hypotheses and related experiments provided below, do the following:

A) *Determine the null hypothesis.*

B) *Make predictions regarding the results of the experiment based on the alternate hypothesis and on the null hypothesis. What would the response variable(s) be for each experiment?*

C) *Make graphs representing your predicted results. What would go on the axes of your graphs? Which of the variables are the dependent and independent variables? You might not be able to specify the exact values predicted for the data for each hypothesis, but you should be able to depict how they would compare on the graph.*

Sample hypotheses with brief descriptions of experiments

1. *Acetylcholine stimulates muscle contraction.*

Cultures of muscle fibers are prepared. A solution of acetylcholine is applied to some of the cultures at different concentrations. Other cultures receive only the solvent without acetylcholine, and still others have nothing applied to them.

2. *Water moves from the roots of a plant upwards because of transpiration, where evaporation of water from leaves draws water up through the xylem.*

Plants are placed in a medium where their roots are submerged in water with a dye. Some plants have their stomata, tiny openings on the undersides of the leaves that allow for gas exchange, painted shut with nail polish. Other plants have some leaves painted, others not painted. Still other plants are not painted at all.

3. *Amphibian metamorphosis is regulated by the thyroid hormone thyroxin.*

Tadpoles are kept in a number of tanks. Some tanks have the thyroid hormone thyroxin added at different concentrations. Other tanks have no thyroxin added.

1.4 Tests and their design

We have discussed how hypotheses can lead to predictions, and comparing predictions to observations is ultimately how we test hypotheses. If the battery is dead, we predict that the radio will not work, and if the battery is not dead the radio will work. Those predictions provided a simple test of our hypothesis. Most tests of scientific hypotheses are more complex. Hypotheses make many predictions, but which predictions will be most helpful for a test? What other things could produce the observations that we predict for our hypothesis, and how can we discriminate between causes related to our hypothesis and other kinds of causes? What do we measure and how do we measure it in order to know whether or not our predictions have been met? Even if we can measure something, how do we know if it has changed in the way we predicted or not? The answers to these questions are all part of designing a good test of a hypothesis.

We typically call these well-designed tests **experiments**, but note that there are different kinds of experiments. The word "experiment" often conjures up images of antiseptic laboratories, glassware with graduated measurements, and people in white coats. Certainly, many experiments are performed in labs, because labs provide environments amenable to good experiments. But experiments can—and sometimes must—occur outside of the lab.

Good experiments have a number of qualities. First and foremost, the results of the experiment will be different depending on whether one hypothesis or another is true. In other words, in a good experiment, the null hypothesis will predict different results than what is predicted by the alternate hypothesis. If the hypotheses make the same predictions, then the test will not allow us to determine whether one hypothesis fits the data better than the other. We can then ask whether the actual results match the predictions of one hypothesis or the other, or perhaps even of neither.

Good tests also provide a way to minimize the effects of factors other than the one about which we are making predictions. This is the function of a **controlled experiment**, where we control all of the relevant variables in an experiment, whether they pertain to the cause we are testing or to other factors. Controlled experiments also typically include **control treatments**. Control treatments differ from **experimental treatments** in that they have not been manipulated in terms of the variable that is being tested. For instance, in our car example, we could control for other causes of a radio not working. We could start the engine and turn on the radio of another car, which would allow us to know what we should hear on the radio if it is working. We could replace the car's battery with a new one (or one known to be functioning) and see if the radio is working.

As another example, let's say you wanted to test whether a particular drug has a particular effect on mice. You could administer the drug to mice in your experimental treatments but not to those in your control treatments, and all mice in all treatments would otherwise be as similar as possible. This way, if the mice in the experimental treatments exhibit effects that differ from those in the control treatments, we can have the greatest confidence that those differences are due to the effect of the drug.

In essence, the control treatments give us a point of comparison for our experimental treatments. In the plant experiment, our experimental treatments were the plants we placed in the dark. We predict that these plants would not grow as well as our control plants, which were the ones in the sunlight. Without the control plants, we would have no way of evaluating whether what happened to the plants in the dark was really affected by the lack of sunlight.

Activity

Using the three hypotheses and experiments in the previous activity, do the following:

Identify the control treatment(s).

Use the results given in the tables below to determine whether the data were consistent with the predictions of the alternate or null hypotheses.

1. *Acetylcholine experiment*

Type of treatment	Percentage contracting
Acetylcholine added	100
Solvent only	0
No acetylcholine or solvent	0

2. *Water movement in plants*

Type of treatment	Movement of water (as percentage of plant height)
All stomata covered	0
Half of stomata covered	50
No stomata covered	100

3. *Amphibian metamorphosis*

Type of treatment	Percentage metamorphosing
Thyroxin added	85
No thyroxin added	2

There are additional qualities of a good experiment. Good experiments include more than one subject, to ensure that the results are not biased by the peculiarities of a single individual; the number of subjects is often what we refer to as the **sample size**. Most experiments strive to have a large enough sample size to make comparisons meaningful and to enhance the utility of any statistics that are applied to the data.

Good experiments also account for other **confounding variables** that could be affecting the results, particularly those pertaining to **environmental variation** and **individual variation**. Environmental variation includes a number of conditions pertaining to the

environment in which the experiment is run. These would include things like temperature, air pressure, and lighting conditions, but also things like the nutrition and habitats of living subjects, the timing of treatments, and other conditions that are external to the subjects themselves. Ideally, these environmental conditions will be the same for all subjects in the experiment; otherwise, we might not be able to eliminate the possibility that a difference in the results obtained from different subjects is due to differences in their environments. This is why experiments are often performed in labs, where the environment is carefully controlled and maintained to be as identical as possible for all treatments.

Individual subjects are not identical, so how do we determine whether a difference between the results from two subjects is not due to the fact that they are simply different individuals? Good experiments therefore need to control for individual variation. For instance, many experiments use model organisms, such as specially bred strains of mice or bacteria, where all individuals are as close to being genetically identical as they can be. Even if your mice are genetically identical, they might still be different due to the different lives that they have led. Thus, you would also want to make sure that your mice were of similar size, age, and health.

In many cases, you cannot simply select subjects that are nearly identical. This is especially true for clinical studies of human health. How does one control for individual variation then? In such cases, a researcher can design an experiment to minimize **bias**, or some kind of nonrandom variation in the treatments. For instance, if your control mice were all males, but your experimental mice were all females, that would be a very biased distribution of sexes in your treatments, and you would not be able to exclude the possibility that the differences in the results between your controls and experimentals were due to sex differences. In experiments like clinical trials, bias is avoided by **randomization**; subjects are assigned randomly to a treatment group, and with large enough sample sizes each treatment should have a set of individuals who, while not identical to each other, vary in roughly the same way from treatment group to treatment group.

Finally, good tests are or can be repeated, by the original researcher or by those in another lab. The terms **repetition** and **replication** are often used loosely for either kind of repeating an experiment, but typically repetition refers to essentially having multiple instances or runs of an experiment when it is carried out, whereas replication typically refers to someone else carrying out the same experiment to see if they get similar results. Note that, in this case, repetition might mean running an experiment several times in sequence, or having multiple control and treatment groups. In many cases, simply using multiple subjects is a form of repetition.

Activity

A) *For the three experiments in the previous activities, identify the additional information you would need if you were going to run these experiments yourself.*
B) *What would you need to do in each of the experiments to execute them and to control for confounding variables?*
C) *How many individuals would you need, and how would you take into account the effects of differences between individuals?*

A Closer Look: Descriptive Statistics and Calculating Means and Standard Deviations

An important part of experimental design is determining what you will measure and how you will measure it. After you make your measurements, how will you interpret them? For instance, in the plant experiment, when we measure the heights of our plants, it is very likely that they will not all be the same height, but how different should they be for us to say whether the prediction of a hypothesis is matched or not? If we use many plants in each treatment, we will need a way to describe the growth of each treatment as a whole. Descriptive statistics help us to do that. Two of the most basic descriptive statistics are the mean and the standard deviation. The **mean** is the average value of a measurement (say, of a particular trait) for all of the sampled individuals. While the mean is helpful, it doesn't provide a very complete picture of variation in this trait in the population. In other words, we might glean something from the fact that the mean is large or small, but is that mean derived from a sample that varies very little from the mean or that has a wide range of variation in this trait? A useful statistic for understanding the variation of a trait among sampled individuals around its mean is the **standard deviation**. The standard deviation is essentially the average amount by which individuals vary from the mean. If the range of variation is wide, then the standard deviation will be large relative to the mean; if there is little variation in the sample, then the standard deviation will be small relative to the mean.

The formulas below are for calculating means and standard deviations.

Mean: $\overline{X} = \frac{\sum x_i}{n}$ Standard deviation: $\sqrt{\frac{\sum (x_i - \overline{X})^2}{n-1}}$

The mean is fairly straightforward and probably familiar to most students, even if its representation in the above formula is not. The mean takes the sum of all the individual values for a given variable (in this case, x, with each individual's value for x given as x_i); that sum is Σx_i in the formula. We then divide that sum by the number of individuals in the sample, or n.

The formula for standard deviation shows some similarities to the formula for the mean: there is a sum being made that involves x_i, and we are dividing by a number close to the sample size ($n-1$). The main differences are that (a) we are subtracting the mean from x_i, (b) we're squaring that difference, and (c) we're taking the square root of the whole calculation. The first difference should make sense: we are interested in how individuals vary from the mean, on average, so we can start by calculating the difference between each individual's value for x and the mean. But why square this difference? The reason is that individuals vary both by being less than the mean and by being greater than the mean. Thus, if we just take differences from the mean, we will end up with some positive numbers and some negative numbers, and the sum of those will be zero. By squaring the difference, we eliminate the sign, and all squared differences will be positive. Thus, the resulting sum will be positive. By taking the square root, we are essentially "undoing" that squaring and ending up with a number that reflects how the population tends to vary from the mean.

The standard deviation is useful for describing how much a sample varies. Here is a simple example to illustrate this.

1.5 Extending tests beyond the lab

Typically, we think of tests as **manipulative** experiments; in other words, the scientist manipulates the conditions of the experiment, creating a situation that does not exist outside of the lab, and which allows for the greatest control of confounding variables. Indeed, the controlled experiment is the standard for testing in science. In some cases, controlled experiments are not feasible, but we may still be able to test hypotheses through **observational experiments**, using comparative methods and "natural experiments." For instance, if we wanted to test the hypothesis that the number of species on an island is determined by the size (area) of the island, we could compare the number of species on lots of islands of different sizes. Although we can't create a control treatment for such a study, we can still try to control for confounding variables by how we sample islands. For instance, we can restrict our sample to islands within a certain range of latitudes, in order to control for the effect of latitude.

As another example, consider the three-spined stickleback (*Gasterosteus aculeatus*; Figure 1.1). These fish are found in marine and freshwater environments in the northern part of the Northern Hemisphere; marine populations have managed to establish themselves in inland lakes and streams numerous times. In the marine populations, the fish are "armored" with bony plates and spines along their flanks and undersides. Some freshwater populations retain this armor, but in others this armor is greatly reduced or even absent. One hypothesis is that the armor, which costs the fish considerable energy to make, is an adaptation for protection against predators. Based on this hypothesis, we would predict that sticklebacks would be well-armored when populations are in places with many predators. If we looked at different lakes with sticklebacks, we would predict that we'd find armored fish in the lakes with predators and unarmored fish in those that lack predators. We would not be manipulating the lakes or the fish populations, but instead we would rely on the natural occurrence of lakes that have sticklebacks but that differ in the presence or absence of predators.

Jack perks/Shutterstock.com

Figure 1.1. A three-spined stickleback (*Gasterosteus aculeatus*).

Observational experiments obviously lack some of the opportunities for controlling and manipulating conditions that manipulative experiments enjoy, but that is not to say that observational experiments cannot be as powerful, and they are often necessary. Nor are observational experiments only typical of studies involving observations of plants and animals in the field. Much of what we know about astronomy comes from careful observation of the night sky. Many studies important for our understanding of human health are actually observational studies. A prominent example involves the investigation of the alleged relationship between the measles-mumps-rubella (MMR) vaccine and autism in children. Several observational studies in Finland, Denmark, and the United States examined vaccination and hospital records looking for correlations between autism and vaccination, including the timing of both. In some cases, the researchers compared the incidence of autism in children who had been vaccinated with that of children who had not been vaccinated; the latter group acted as a control group. No significant difference was found in the incidence of autism in children who had and who had not been vaccinated, and there was no relationship between when children were vaccinated and the development of autism. Despite the fact that this experiment could not be carried out in a laboratory, it still provides conclusive evidence regarding the lack of a causal relationship between vaccines and autism.

1.6 Where do we go from here?

An elegant test might allow us to reject competitors and settle on a particular hypothesis. Regardless, while a hypothesis can be falsified, it can be supported but can't be deductively proved to be absolutely true, so there may still be room for more testing. Other scientists may repeat a test to see if they can corroborate the results. More often, answering one question leads to other questions, and the process begins again.