CHAPTER 6

Designing Successful Field Studies

The proper analysis of data goes hand in hand with an appropriate sampling design and experimental layout. If there are serious errors or problems in the design of the study or in the collection of the data, rarely is it possible to repair these problems after the fact. In contrast, if the study is properly designed and executed, the data can often be analyzed in several different ways to answer different questions. In this chapter, we discuss the broad issues that you need to consider when designing an ecological study. We can't overemphasize the importance of thinking about these issues *before* you begin to collect data.

What Is the Point of the Study?

Although it may seem facetious and the answer self-evident, many studies are initiated without a clear answer to this central question. Most answers will take the form of a more focused question.

Are There Spatial or Temporal Differences in Variable *Y*?

This is the most common question that is addressed with survey data, and it represents the starting point of many ecological studies. Standard statistical methods such as analysis of variance (ANOVA) and regression are well-suited to answer this question. Moreover, the conventional testing and rejection of a simple null hypothesis (see Chapter 4) yields a dichotomous yes/no answer to this question. It is difficult to even discuss mechanisms without some sense of the spatial or temporal pattern in your data. Understanding the forces controlling biological diversity, for example, requires at a minimum a spatial map of species richness. The design and implementation of a successful ecological survey requires a great deal of effort and care, just as much as is needed for a successful experimental study. In some cases, the survey study will address all of your research goals; in other cases, a survey study will be the first step in a research

project. Once you have documented spatial and temporal patterns in your data, you will conduct experiments or collect additional data to address the mechanisms responsible for those patterns.

What Is the Effect of Factor *X* on Variable *Y*?

This is the question directly answered by a manipulative experiment. In a field or laboratory experiment, the investigator actively establishes different levels of Factor *X* and measures the response of Variable *Y*. If the experimental design and statistical analysis are appropriate, the resulting *P*-value can be used to test the null hypothesis of no effect of Factor *X*. Statistically significant results suggest that Factor *X* influences Variable *Y*, and that the "signal" of Factor *X* is strong enough to be detected above the "noise" caused by other sources of natural variation.[1] Certain natural experiments can be analyzed in the same way, taking advantage of natural variation that exists in Factor *X*. However, the resulting inferences are usually weaker because there is less control over confounding variables. We discuss natural experiments in more detail later in this chapter.

Are the Measurements of Variable *Y* Consistent with the Predictions of Hypothesis H?

This question represents the classic confrontation between theory and data (Hilborn and Mangel 1997). In Chapter 4, we discussed two strategies we use for this confrontation: the inductive approach, in which a single hypothesis is recursively modified to conform to accumulating data, and the hypothetico-deductive approach, in which hypotheses are falsified and discarded if they do not predict the data. Data from either experimental or observational studies can be used to ask whether observations are consistent with the predictions of a mechanistic hypothesis. Unfortunately, ecologists do not always state this question so plainly. Two limitations are (1) many ecological hypotheses do not generate simple, falsifiable predictions; and (2) even when an hypothesis does generate predictions, they are rarely unique. Therefore, it may not be possible to definitively test Hypothesis H using only data collected on Variable *Y*.

[1] Although manipulative experiments allow for strong inferences, they may not reveal explicit mechanisms. Many ecological experiments are simple "black box" experiments that measure the *response* of the Variable *Y* to changes in Factor *X*, but do not elucidate lower-level mechanisms causing that response. Such a mechanistic understanding may require additional observations or experiments addressing a more focused question about process.

Using the Measurements of Variable *Y*, What Is the Best Estimate of Parameter θ in Model *Z*?

Statistical and mathematical models are powerful tools in ecology and environmental science. They allow us to forecast how populations and communities will change through time or respond to altered environmental conditions (e.g., Sjögren-Gulve and Ebenhard 2000). Models can also help us to understand how different ecological mechanisms interact simultaneously to control the structure of communities and populations (Caswell 1988). Parameter estimation is required for building predictive models and is an especially important feature of Bayesian analysis (see Chapter 5). Rarely is there a simple one-to-one correspondence between the value of Variable *Y* measured in the field and the value of Parameter θ in our model. Instead, those parameters have to be extracted and estimated indirectly from our data. Unfortunately, some of the most common and traditional designs used in ecological experiments and field surveys, such as the analysis of variance (see Chapter 10), are not very useful for estimating model parameters. Chapter 7 discusses some alternative designs that are more useful for parameter estimation.

Manipulative Experiments

In a **manipulative experiment**, the investigator first alters levels of the predictor variable (or factor), and then measures how one or more variables of interest respond to these alterations. These results are then used to test hypotheses of cause and effect. For example, if we are interested in testing the hypothesis that lizard predation controls spider density on small Caribbean islands, we could alter the density of lizards in a series of enclosures and measure the resulting density of spiders (e.g., Spiller and Schoener 1998). We could then plot these data in a graph in which the *x*-axis (= independent variable) is lizard density, and the *y*-axis (= dependent variable) is spider density (Figure 6.1A,B).

Our null hypothesis is that there is no relationship between these two variables (Figure 6.1A). That is, spider density might be high or low in a particular enclosure, but it is not related to the density of lizards that were established in the enclosure. Alternatively, we might observe a negative relationship between spider and lizard density: enclosures with the highest lizard density have the fewest spiders, and vice-versa (Figure 6.1B). This pattern then would have to be subject to a statistical analysis such as regression (see Chapter 9) to determine whether or not the evidence was sufficient to reject the null hypothesis of no relationship between lizard and spider densities. From these data we could also estimate regression model parameters that quantify the strength of the relationship.

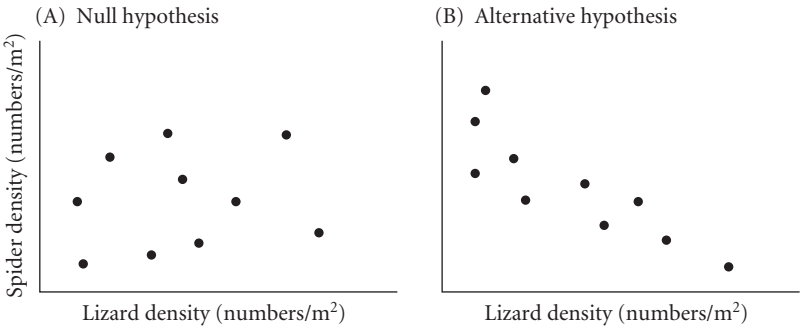

Figure 6.1 Relationship between lizard density and spider density in manipulative and natural field experiments. Each point represents a plot or quadrat in which both spider density and lizard density have been measured. (A) The null hypothesis is that lizard density has no effect on spider density. (B) The alternative hypothesis is that lizard predation controls spider density, leading to a negative relationship between these two variables.

Although field experiments are popular and powerful, they have several important limitations. First, it is challenging to conduct experiments on large spatial scales; over 80% of field experiments have been conducted in plots of less than 1 m^2 (Kareiva and Anderson 1988; Wiens 1989). When experiments are conducted on large spatial scales, replication is inevitably sacrificed (Carpenter 1989). Even when they are properly replicated, experiments conducted on small spatial scales may not yield results that are representative of patterns and processes occurring at larger spatial scales (Englund and Cooper 2003).

Second, field experiments are often restricted to relatively small-bodied and short-lived organisms that are easy to manipulate. Although we always want to generalize the results of our experiments to other systems, it is unlikely that the interaction between lizards and spiders will tell us much about the interaction between lions and wildebeest. Third, it is difficult to change one and only one variable at a time in a manipulative experiment. For example, cages can exclude other kinds of predators and prey, and introduce shading. If we carelessly compare spider densities in caged plots versus uncaged "controls," the effects of lizard predation are **confounded** with other physical differences among the treatments. We discuss solutions to confounding variables later in this chapter.

Finally, many standard experimental designs are simply unwieldy for realistic field experiments. For example, suppose we are interested in investigating competitive interactions in a group of eight spider species. Each treatment in

such an experiment would consist of a unique combination of species. Although the number of species in each treatment ranges from only 1 to 8, the number of unique combinations is $2^8 - 1 = 255$. If we want to establish even 10 replicates of each treatment (see "The Rule of 10," discussed later in this chapter), we need 2550 plots. That may not be possible because of constraints on space, time, or labor. Because of all these potential limitations, many important questions in community ecology cannot be addressed with field experiments.

Natural Experiments

A **natural experiment** (Cody 1974) is not really an experiment at all. Instead, it is an observational study in which we take advantage of natural variation that is present in the variable of interest. For example, rather than manipulate lizard densities directly (a difficult, expensive, and time-consuming endeavor), we could census a set of plots (or islands) that vary naturally in their density of lizards (Schoener 1991). Ideally, these plots would vary *only* in the density of lizards and would be identical in all other ways. We could then analyze the relationship between spider density and lizard density as illustrated in Figure 6.1.

Natural experiments and manipulative experiments superficially generate the same kinds of data and are often analyzed with the same kinds of statistics. However, there are often important differences in the interpretation of natural and manipulative experiments. In a manipulative experiment, if we have established valid controls and maintained the same environmental conditions among the replicates, any consistent differences in the response variable (e.g., spider density) can be attributed confidently to differences in the manipulated factor (e.g., lizard density).

We don't have this same confidence in interpreting results of natural experiments. In a natural experiment, we do not know the direction of cause and effect, and we have not controlled for other variables that surely will differ among the replicates. For the lizard–spider example, there are at least four hypotheses that could account for a negative association between lizard and spider densities:

1. Lizards may control spider density. This was the alternative hypothesis of interest in the original field experiment.
2. Spiders may directly or indirectly control lizard density. Suppose, for example, that large hunting spiders consume small lizards, or that spiders are also preyed upon by birds that feed on lizards. In both cases, increasing spider density may decrease lizard density, even though lizards do feed on spiders.

3. Both spider and lizard densities are controlled by an unmeasured environmental factor. For example, suppose that spider densities are highest in wet plots and lizard densities are highest in dry plots. Even if lizards have little effect on spiders, the pattern in Figure 6.1B will emerge: wet plots will have many spiders and few lizards, and dry plots will have many lizards and few spiders.
4. Environmental factors may control the strength of the interaction between lizards and spiders. For example, lizards might be efficient predators on spiders in dry plots, but inefficient predators in wet plots. In such cases, the density of spiders will depend on both the density of lizards and the level of moisture in the plot (Spiller and Schoener 1995).

These four scenarios are only the simplest ones that might lead to a negative relationship between lizard density and spider density (Figure 6.2). If we add double-headed arrows to these diagrams (lizards and spiders reciprocally affect one another's densities), there is an even larger suite of hypotheses that could account for the observed relationships between spider density and lizard density (see Figure 6.1).

All of this does not mean that natural experiments are hopeless, however. In many cases we can collect additional data to distinguish among these hypotheses. For example, if we suspect that environmental variables such as moisture

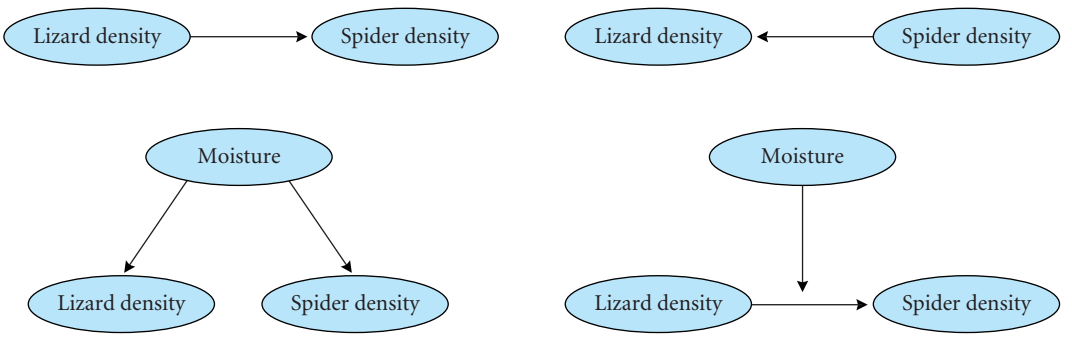

Figure 6.2 Mechanistic hypotheses to account for correlations between lizard density and spider density (see Figure 6.1). The cause-and-effect relationship might be from predator to prey (upper left) or prey to predator (upper right). More complicated models include the effects of other biotic or abiotic variables. For example, there might be no interaction between spiders and lizards, but densities of both are controlled by a third variable, such as moisture (lower left). Alternatively, moisture might have an indirect effect by altering the interaction of lizards and spiders (lower right).

are important, we either can restrict the survey to a set of plots with comparable moisture levels, or (better still) measure lizard density, spider density, and moisture levels in a series of plots censused over a moisture gradient. Confounding variables and alternative mechanisms also can be problematic in field experiments. However, their impacts will be reduced if the investigator conducts the experiment at an appropriate spatial and temporal scale, establishes proper controls, replicates adequately, and uses randomization to locate replicates and assign treatments.

Overall, manipulative experiments allow for greater confidence in our inferences about cause and effect, but they are limited to relatively small spatial scales and short time frames. Natural experiments can be conducted at virtually any spatial scale (small quadrats to entire continents) and over any time interval (weekly field measurements, to annual censuses, to fossil strata). However, it is more challenging to tease apart cause-and-effect relationships in natural experiments.[2]

Snapshot versus Trajectory Experiments

Two variants of the natural experiment are the **snapshot experiment** and the **trajectory experiment** (Diamond 1986). Snapshot experiments are replicated in space, and trajectory experiments are replicated in time. For the data in Figure 6.1, suppose we censused 10 different plots in a single day. This is a snapshot experiment in which the replication is spatial; each observation represents a different plot censused at the same time. On the other hand, suppose we visited a single plot in 10 different years. This is a trajectory experiment in which the replication is temporal; each observation represents a different year in the study.

The advantages of a snapshot experiment are that it is rapid, and the spatial replicates arguably are more statistically independent of one another than are

[2] In some cases, the distinction between manipulative and natural field experiments is not clear-cut. Human activity has generated many unintended large-scale experiments including eutrophication, habitat alteration, global climate change, and species introductions and removals. Imaginative ecologists can take advantage of these alterations to design studies in which the confidence in the conclusions is very high. For example, Knapp et al. (2001) studied the impacts of trout introductions to lakes in the Sierra Nevada by comparing invertebrate communities in naturally fishless lakes, stocked lakes, and lakes that formerly were stocked with fish. Many comparisons of this kind are possible to document consequences of human activity. However, as human impacts become more widespread and pervasive, it may be harder and harder to find sites that can be considered unmanipulated controls.

the temporal replicates of a trajectory experiment. The majority of ecological data sets are snapshot experiments, reflecting the 3- to 5-year time frame of most research grants and dissertation studies.[3] In fact, many studies of temporal change are actually snapshot studies, because variation in space is treated as a proxy variable for variation in time. For example, successional change in plant communities can be studied by sampling from a chronosequence—a set of observations, sites, or habitats along a spatial gradient that differ in the time of origin (e.g., Law et al. 2003).

The advantage of a trajectory experiment is that it reveals how ecological systems change through time. Many ecological and environmental models describe precisely this kind of change, and trajectory experiments allow for stronger comparisons between model predictions and field data. Moreover, many models for conservation and environmental forecasting are designed to predict future conditions, and data for these models are derived most reliably from trajectory experiments. Many of the most valuable data sets in ecology are long time-series data for which populations and communities at a site are sampled year after year with consistent, standardized methods. However, trajectory experiments that are restricted to a single site are unreplicated in space. We don't know if the temporal trajectories described from that site are typical for what we might find at other sites. Each trajectory is essentially a sample size of *one* at a given site.[4]

The Problem of Temporal Dependence

A more difficult problem with trajectory experiments is the potential non-independence of data collected in a temporal sequence. For example, suppose you measure tree diameters each month for one year in a plot of redwood trees. Red-

[3] A notable exception to short-term ecological experiments is the coordinated set of studies developed at Long Term Ecological Research (LTER) sites. The National Science Foundation (NSF) funded the establishment of these sites throughout the 1980s and 1990s specifically to address the need for ecological research studies that span decades to centuries. See www.lternet.edu/.

[4] Snapshot and trajectory designs show up in manipulative experiments as well. In particular, some designs include a series of measurements taken before and after a manipulation. The "before" measurements serve as a type of "control" that can be compared to the measurements taken after the manipulation or intervention. This sort of BACI design (Before-After, Control-Impact) is especially important in environmental impact analysis and in studies where spatial replication may be limited. For more on BACI, see the section "Large Scale Studies and Environmental Impacts" later in this chapter, and see Chapter 7.

woods grow very slowly, so the measurements from one month to the next will be virtually identical. Most foresters would say that you don't have 12 independent data points, you have only one (the average diameter for that year). On the other hand, monthly measurements of a rapidly developing freshwater plankton community reasonably could be viewed as statistically independent of one another. Naturally, the further apart in time the samples are separated from one another, the more they function as independent replicates.

But even if the correct census interval is used, there is still a subtle problem in how temporal change should be modeled. For example, suppose you are trying to model changes in population size of a desert annual plant for which you have access to a nice trajectory study, with 100 years of consecutive annual censuses. You could fit a standard linear regression model (see Chapter 9) to the time series

$$N_t = \beta_0 + \beta_1 t + \varepsilon \tag{6.1}$$

In this equation, population size (N_t) is a linear function of the amount of time (t) that has passed. The coefficients β_0 and β_1 are the intercept and slope of this straight line. If β_1 is less than 0.0, the population is shrinking with time, and if $\beta_1 > 0$, N is increasing. Here ε is a normally distributed **white noise**[5] error term that incorporates both measurement error and random variation in population size. Chapter 9 will explain this model in much greater detail, but we introduce it now as a simple way to think about how population size might change in a linear fashion with the passage of time.

However, this model does not take into account that population size changes through births and deaths affecting *current* population size. A **time-series model** would describe population growth as

$$N_{t+1} = \beta_0 + \beta_1 N_t + \varepsilon \tag{6.2}$$

[5] White noise is a type of error distribution in which the errors are independent and uncorrelated with one another. It is called white noise as an analogy to white light, which is an equal mixture of short and long wavelengths. In contrast, red noise is dominated by low-frequency perturbations, just as red light is dominated by low-frequency light waves. Most time series of population sizes exhibit a reddened noise spectrum (Pimm and Redfearn 1988), so that variances in population size increase when they are analyzed at larger temporal scales. Parametric regression models require normally distributed error terms, so white noise distributions form the basis for most stochastic ecological models. However, an entire spectrum of colored noise distributions (1/*f* noise) may provide a better fit to many ecological and evolutionary datasets (Halley 1996).

In this model, the population size in the next time step (N_{t+1}) depends not simply on the amount of time t that has passed, but rather on the population size at the last time step (N_t). In this model, the constant β_1 is a multiplier term that determines whether the population is exponentially increasing ($\beta_1 > 1.0$) or decreasing ($\beta_1 < 1.0$). As before, ε is a white noise error term.

The linear model (Equation 6.1) describes a simple *additive* increase of N with time, whereas the time-series, or **autoregressive** model (Equation 6.2) describes an *exponential* increase, because the factor β_1 is a multiplier that, on average, gives a constant percentage increase in population size at each time step. The more important difference between the two models, however, is that the differences between the observed and predicted population sizes (i.e., the **deviations**) in the time-series model are correlated with one another. As a consequence, there tend to be runs, or periods of consecutive increases followed by periods of consecutive decreases. This is because the growth trajectory has a "memory"—each consecutive observation (N_{t+1}) depends directly on the one that came before it (the N_t term in Equation 6.2). In contrast, the linear model has no memory, and the increases are a function only of time (and ε), and not of N_t. Hence, the positive and negative deviations follow one another in a purely random fashion (Figure 6.3). Correlated deviations, which are typical of data collected in trajectory studies, violate the assumptions of most conventional statistical analyses.[6] Analytical and computer-intensive methods have been developed for analyzing both sample data and experimental data collected through time (Ives et al. 2003; Turchin 2003).

This does not mean we cannot incorporate time-series data into conventional statistical analyses. In Chapters 7 and 10, we will discuss additional ways to analyze time-series data. These methods require that you pay careful attention to both the sampling design and the treatment of the data after you have collected them. In this respect, time-series or trajectory data are just like any other data.

Press versus Pulse Experiments

In manipulative studies, we also distinguish between **press experiments** and **pulse experiments** (Bender et al. 1984). In a press experiment, the altered conditions in the treatment are maintained through time and are re-applied as necessary to ensure that the strength of the manipulation remains constant. Thus,

[6] Actually, spatial autocorrelation generates the same problems (Legendre and Legendre 1998; Lichstein et al. 2003). However, tools for spatial autocorrelation analysis have developed more or less independently of time-series analyses, perhaps because we perceive time as a strictly one-dimensional variable and space as a two- or three-dimensional variable.

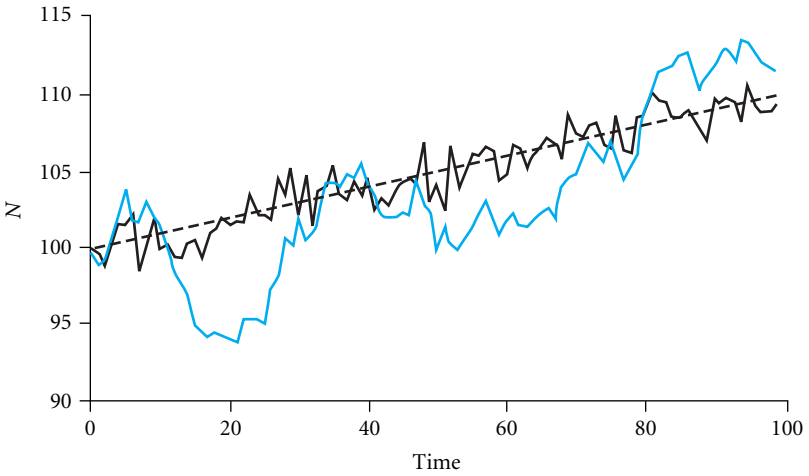

Figure 6.3 Examples of deterministic and stochastic time series, with and without autocorrelation. Each population begins with 100 individuals. A linear model without error (dashed line) illustrates a constant upward trend in population data. A linear model with a stochastic white noise error term (black line) adds temporally uncorrelated variability. Finally, an autocorrelated model (blue line) describes population size in the next time step ($t + 1$) as a function of the population size in the current time step (t) plus random noise. Although the error term in this model is still a simple random variable, the resulting time series shows autocorrelation—there are runs of population increases followed by runs of population decreases. For the linear model and the stochastic white noise model, the equation is $N_t = a + bt + \varepsilon$, with $a = 100$ and $b = 0.10$. For the autocorrelated model, $N_{t+1} = a + bN_t + \varepsilon$, with $a = 0.0$ and $b = 1.0015$. For both models with error, ε is a normal random variable: $\varepsilon \sim N(0,1)$.

fertilizer may have to be re-applied to plants, and animals that have died or disappeared from a plot may have to be replaced. In contrast, in a pulse experiment, experimental treatments are applied only once, at the start of the study. The treatment is not re-applied, and the replicate is allowed to "recover" from the manipulation (Figure 6.4A).

Press and pulse experiments measure two different responses to the treatment. The press experiment (Figure 6.4B) measures the resistance of the system to the experimental treatment: the extent to which it resists change in the constant environment created by the press experiment. A system with low resistance will exhibit a large response in a press experiment, whereas a system with high resistance will exhibit little difference between control and manipulated treatments.

The pulse experiment measures the resilience of the system to the experimental treatment: the extent to which the system recovers from a single perturbation. A system with high resilience will show a rapid return to control con-

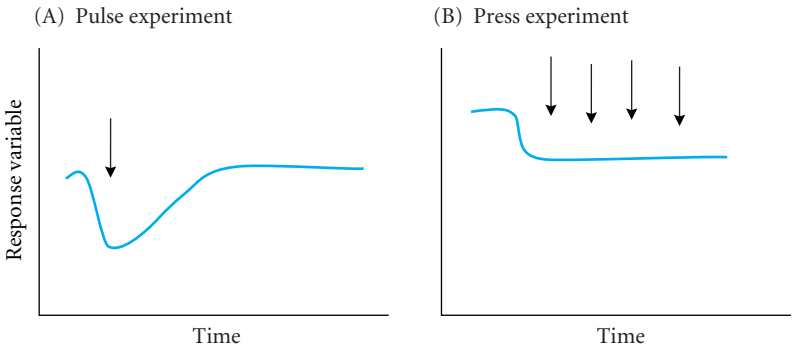

Figure 6.4 Ecological pulse and press experiments. The arrow indicates a treatment application, and the line indicates the temporal trajectory of the response variable. The pulse experiment (A) measures the response to a single treatment application (resilience), whereas the press experiment (B) measures the response under constant conditions (resistance).

ditions, whereas a system with low resilience will take a long time to recover; control and manipulated plots will continue to differ for a long time after the single treatment application.

The distinction between press and pulse experiments is not in the number of treatment applications used, but in whether the altered conditions are maintained through time in the treatment. If environmental conditions remain constant following a single perturbation for the duration of the experiment, the design is effectively a press experiment. Another distinction between press and pulse experiments is that the press experiment measures the response of the system under equilibrium conditions, whereas the pulse experiment records transient responses in a changing environment.

Replication

How Much Replication?

This is one of the most common questions that ecologists and environmental scientists ask of statisticians. The correct response is that the answer depends on the variance in the data and the **effect size**—the difference that you wish to detect between the averages of the groups being compared. Unfortunately, these two quantities may be difficult to estimate, although you always should consider what effect size would be reasonable to observe.

To estimate variances, many statisticians will recommend that you conduct a pilot study. Unfortunately, pilot studies usually are not feasible—you rarely have the freedom to set up and run a costly or lengthy study more than once.

Field seasons and grant proposals are too short for this sort of luxury. However, you may be able to estimate reasonable ranges of variances and effect sizes from previously published studies and from discussions with colleagues. You can then use these values to determine the statistical power (see Chapter 4) that will result from different combinations of replicates, variances, and effect sizes (see Figure 4.5 for an example). At a minimum, however, you need to first answer the following question:

How Many Total Replicates Are Affordable?

It takes time, labor, and money to collect either experimental or survey data, and you need to determine precisely the total sample size that you can afford. If you are conducting expensive tissue or sample analyses, the dollar cost may be the limiting factor. However, in many studies, time and labor are more limiting than money. This is especially true for geographical surveys conducted over large spatial scales, for which you (and your field crew if you are lucky enough to have one) may spend as much time traveling to study sites as you do collecting field data. Ideally, all of the replicates should be measured simultaneously, giving you a perfect snapshot experiment. The more time it takes to collect all the data, the more conditions will have changed from the first sample to the last. For experimental studies, if the data are not collected all at once, then the amount of time that has passed since treatment application is no longer identical for all replicates.

Obviously, the larger the spatial scale of the study, the harder it is to collect all of the data within a reasonable time frame. Nevertheless, the payoff may be greater because the scope of inference is tied to the spatial scale of analysis: conclusions based on samples taken only at one site may not be valid at other sites. However, there is no point in developing an unrealistic sampling design. Carefully map out your project from start to finish to ensure it will be feasible.[7] Only once you know the total number of replicates or observations that you can collect can you begin to design your experiment by applying the rule of 10.

[7] It can be very informative to use a stopwatch to time carefully how long it takes to complete a single replicate measurement of your study. Like the efficiency expert father in *Cheaper By The Dozen* (Gilbreth and Carey 1949), we put great stock in such numbers. With these data, we can accurately estimate how many replicates we can take in an hour, and how much total field time we will need to complete the census. The same principle applies to sample processing, measurements that we make back in the laboratory, the entry of data into the computer, and the long-term storage and curation of data (see Chapter 8). All of these activities take time that needs to be accounted for when planning an ecological study.

The Rule of 10

The **Rule of 10** is that you should collect at least 10 replicate observations for each category or treatment level. For example, suppose you have determined that you can collect 50 total observations in a experiment examining photosynthetic rates among different plant species. A good design for a one-way ANOVA would be to compare photosynthetic rates among not more than five species. For each species, you would choose randomly 10 plants and take one measurement from each plant.

The Rule of 10 is not based on any theoretical principle of experimental design or statistical analysis, but is a reflection of our hard-won field experience with designs that have been successful and those that have not. It is certainly possible to analyze data sets with less than 10 observations per treatment, and we ourselves often break the rule. Balanced designs with many treatment combinations but only four or five replicates may be quite powerful. And certain one-way designs with only a few treatment levels may require more than 10 replicates per treatment if variances are large.

Nevertheless, the Rule of 10 is a solid starting point. Even if you set up the design with 10 observations per treatment level, it is unlikely that you will end up with that number. In spite of your best efforts, data may be lost for a variety of reasons, including equipment failures, weather disasters, plot losses, human disturbances or errors, improper data transcription, and environmental alterations. The Rule of 10 at least gives you a fighting chance to collect data with reasonable statistical power for revealing patterns.[8] In Chapter 7, we will discuss efficient sample designs and strategies for maximizing the amount of information you can squeeze out of your data.

Large-Scale Studies and Environmental Impacts

The Rule of 10 is useful for small-scale manipulative studies in which the study units (plots, leaves, etc.) are of manageable size. But it doesn't apply to large-scale ecosystem experiments, such as whole-lake manipulations, because replicates may be unavailable or too expensive. The Rule of 10 also does not apply to many environmental impact studies, where the assessment of an impact is required at a single site. In such cases, the best strategy is to use a **BACI design** (Before-After, Control-Impact). In some BACI designs, the replication is achieved through time:

[8] Another useful rule is the Rule of 5. If you want to estimate the curvature or non-linearity of a response, you need to use at least five levels of the predictor variable. As we will discuss in Chapter 7, a better solution is to use a regression design, in which the predictor variable is continuous, rather than categorical with a fixed number of levels.

the control and impact sites are censused repeatedly both before and after the impact. The lack of spatial replication restricts the inferences to the impact site itself (which may be the point of the study), and requires that the impact is not confounded with other factors that may be co-varying with the impact. The lack of spatial replication in simple BACI designs is controversial (Underwood 1994; Murtaugh 2002b), but in many cases they are the best design option (Stewart-Oaten and Bence 2001), especially if they are used with explicit time-series modeling (Carpenter et al. 1989). We will return to BACI and its alternatives in Chapters 7 and 10.

Ensuring Independence

Most statistical analyses assume that replicates are independent of one another. By **independence**, we mean that the observations collected in one replicate do not have an influence on the observations collected in another replicate. Non-independence is most easily understood in an experimental context. Suppose you are studying the response of hummingbird pollinators to the amount of nectar produced by flowers. You set up two adjacent 5 m × 5 m plots. One plot is a control plot; the adjacent plot is a nectar removal plot in which you drain all of the nectar from the flowers. You measure hummingbird visits to flowers in the two plots. In the control plot, you measure an average of 10 visits/hour, compared to only 5 visits/hour in the removal plot.

However, while collecting the data, you notice that once birds arrive at the removal plot, they immediately leave, and *the same birds* then visit the adjacent control plot (Figure 6.5A). Clearly, the two sets of observations are not independent of one another. If the control and treatment plots had been more widely separated in space, the numbers might have come out differently, and the average in the control plots might have been only 7 visits/hour instead of 10 visits/hour (Figure 6.5B). When the two plots are adjacent to one another, non-independence inflates the difference between them, perhaps leading to a spuriously low *P*-value, and a Type I error (incorrect rejection of a true null hypothesis; see Chapter 4). In other cases, non-independence may decrease the apparent differences between treatments, contributing to a Type II error (incorrect acceptance of a false null hypothesis). Unfortunately, non-independence inflates or deflates both *P*-values and power to unknown degrees.

The best safeguard against non-independence is to ensure that replicates within and among treatments are separated from one another by enough space or time to ensure that the they do not affect one another. Unfortunately, we rarely know what that distance or spacing should be, and this is true for both experimental and observational studies. We should use common sense and as much

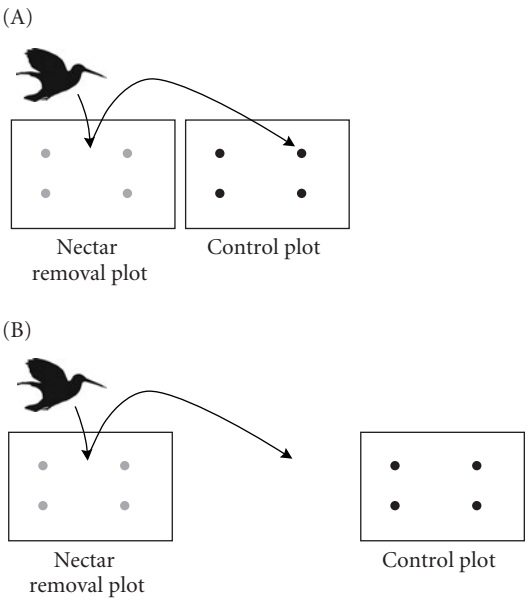

Figure 6.5 The problem of non-independence in ecological studies is illustrated by an experimental design in which hummingbirds forage for nectar in control plots and in plots from which nectar has been removed from all of the flowers. (A) In a non-independent layout, the nectar removal and control plots are adjacent to one another, and hummingbirds that enter the nectar removal plot immediately leave and begin foraging in the adjacent control plot. As a consequence, the data collected in the control plot are not independent of the data collected in the nectar removal plot: the responses in one treatment influence the responses in the other. (B) If the layout is modified so that the two plots are widely separated, hummingbirds that leave the nectar removal plot do not necessarily enter the control plot. The two plots are independent, and the data collected in one plot are not influenced by the presence of the other plot. Although it is easy to illustrate the potential problem of non-independence, in practice it is can be very difficult to know ahead of time the spatial and temporal scales that will ensure statistical independence.

biological knowledge as possible. Try to look at the world from the organism's perspective to think about how far to separate samples. Pilot studies, if feasible, also can suggest appropriate spacing to ensure independence.

So why not just maximize the distance or time between samples? First, as we described earlier, it becomes more expensive to collect data as the distance between samples increases. Second, moving the samples very far apart can introduce new sources of variation because of differences (heterogeneity) within or among habitats. We want our replicates close enough together to ensure we are sampling relatively homogenous or consistent conditions, but far enough apart to ensure that the responses we measure are independent of one another.

In spite of its central importance, the independence problem is almost never discussed explicitly in scientific papers. In the Methods section of a paper, you are likely to read a sentence such as, "We measured 100 randomly selected seedlings growing in full sunlight. Each measured seedling was at least 50 cm from its nearest neighbor." What the authors mean is, "We don't know how far apart the observations would have to have been in order to ensure independence. However, 50 cm seemed like a fair distance for the tiny seedlings we studied. If we had chosen distances greater than 50 cm, we could not have collected all of our data in full sunlight, and some of the seedlings would have been collected in the shade, which obviously would have influenced our results."

Avoiding Confounding Factors

When factors are confounded with one another, their effects cannot be easily disentangled. Let's return to the hummingbird example. Suppose we prudently separated the control and nectar removal plots, but inadvertently placed the removal plot on a sunny hillside and the control plot in a cool valley (Figure 6.6). Hummingbirds forage less frequently in the removal plot (7 visits/hour), and the two plots are now far enough apart that there is no problem of independence. However, hummingbirds naturally tend to avoid foraging in the cool valley, so the foraging rate also is low in these plots (6 visits/hour). Because the treatments are confounded with temperature differences, we cannot tease apart the effects of foraging preferences from those of thermal preferences. In this case, the two forces largely cancel one another, leading to comparable foraging rates in the two plots, although for very different reasons.

This example may seem a bit contrived. Knowing the thermal preferences of hummingbirds, we would not have set up such an experiment. The problem is that there are likely to be unmeasured or unknown variables—even in an apparently homogenous environment—that can have equally strong effects on our experiment. And, if we are conducting a natural experiment, we are stuck with whatever confounding factors are present in the environment. In an observational study of hummingbird foraging, we may not be able to find plots that differ only in their levels of nectar rewards but do not also differ in temperature and other factors known to affect foraging behavior.

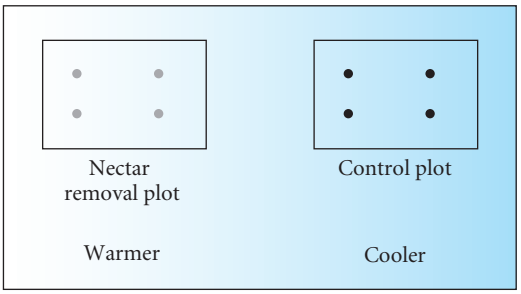

Figure 6.6 A confounded experimental design. As in Figure 6.5, the study establishes control and experimental nectar removal plots to evaluate the responses of foraging hummingbirds. In this design, although the plots are far enough apart to ensure independence, they have been placed at different points along a thermal gradient. Consequently, the treatment effects are confounded with differences in the thermal environment. The net result is that the experiment compares data from a warm nectar removal plot with data from a cool control plot.

Replication and Randomization

The dual threats of confounding factors and non-independence would seem to threaten all of our statistical conclusions and render even our experimental studies suspect. Incorporating **replication** and **randomization** into experimental designs can largely offset the problems introduced by confounding factors and non-independence. By replication, we mean the establishment of multiple plots or observations within the same treatment or comparison group. By randomization, we mean the random assignment of treatments or selection of samples.[9]

Let's return one more time to the hummingbird example. If we follow the principles of randomization and replication, we will set up many replicate control and removal plots (ideally, a minimum of 10 of each). The location of each of these plots in the study area will be random, and the assignment of the treatment (control or removal) to each plot also will be random (Figure 6.7).[10]

How will randomization and replication reduce the problem of confounding factors? Both the warm hillside, the cool valley, and several intermediate sites each will have multiple plots from both control and nectar removal treatments. Thus, the temperature factor is no longer confounded with the treatment, as all treatments occur within each level of temperature. As an additional benefit, this design will also allow you to test the effects of temperature as a covariate on hummingbird foraging behavior, independent of the levels of nectar (see Chapters 7 and 10). It is true that hummingbird visits will still be more frequent on the warm hillside than in the cool valley, but that will be true for replicates of both the control and nectar removal. The temperature will add more variation to the data, but it will not bias the results because the warm and cool plots will

[9] Many samples that are claimed to be random are really **haphazard**. Truly random sampling means using a random number generator (such as the flip of a fair coin, the roll of a fair die, or the use of a reliable computer algorithm for producing random numbers) to decide which replicates to use. In contrast, with haphazard sampling, an ecologist follows a set of general criteria [e.g., mature trees have a diameter of more than 3 cm at breast height (dbh = 1.3 m)] and selects sites or organisms that are spaced homogenously or conveniently within a sample area. Haphazard sampling is often necessary at some level because random sampling is not efficient for many kinds of organisms, especially if their distribution is spatially patchy. However, once a set of organisms or sites is identified, randomization should be used to sample or to assign replicates to different treatment groups.

[10] Randomization takes some time, and you should do as much of it as possible in advance, before you get into the field. It is easy to generate random numbers and simulate random sampling with computer spreadsheets. But it is often the case that you will need to generate random numbers in the field. Coins and dice (especially 10-sided

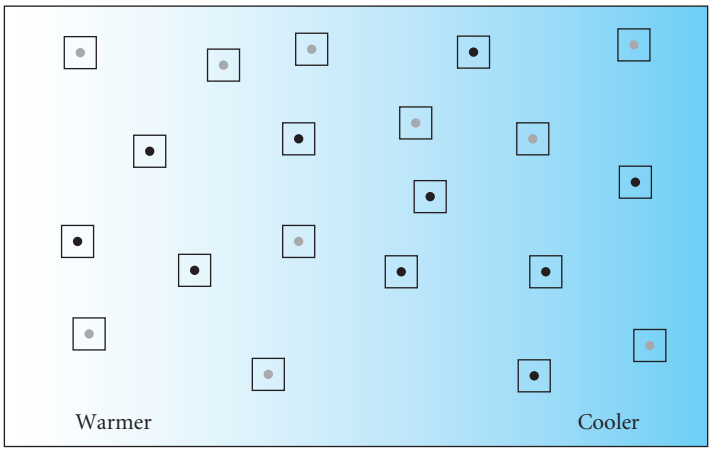

Figure 6.7 A properly replicated and randomized experimental design. The study establishes plots as in Figures 6.6. Each square represents a replicate control plot (black dots) or nectar removal plot (gray dots). The plots are separated by enough distance to ensure independence, but their location within the temperature gradient has been randomized. There are 10 replicates for each of the two treatments. The spatial scale of the drawing is larger than in Figure 6.6.

be distributed approximately equally between the control and removal treatments. Of course, if we knew ahead of time that temperature was an important determinant of foraging behavior, we might not have used this design for the experiment. Randomization minimizes the confounding of treatments with unknown or unmeasured variables in the study area.

It is less obvious how randomization and replication reduce the problem of non-independence among samples. After all, if the plots are too close together, the foraging visits will not be independent, regardless of the amount of replication or randomization. Whenever possible, we should use common sense and

gaming dice) are useful for this purpose. A clever trick is to use a set of coins as a binary random number generator. For example, suppose you have to assign each of your replicates to one of 8 different treatments, and you want to do so randomly. Toss 3 coins, and convert the pattern of heads and tails to a binary number (i.e., a number in base 2). Thus, the first coin indicates the 1s, the second coin indicates the 2s, the third coin indicates the 4s, and so on. Tossing 3 coins will give you a random integer between 0 and 7. If your three tosses are heads, tails, heads (HTH), you have a 1 in the one's place, a 0 in the two's place, and a 1 in the four's place. The number is 1 + 0 + 4 = 5. A toss of (THT) is 0 + 2 + 0 = 2. Three tails gives you a 0 (0 + 0 + 0) and three heads give you a 7 (1 + 2 + 4). Tossing 4 coins will give you 16 integers, and 5 coins will give you 32.

An even easier method is to take a digital stopwatch into the field. Let the watch run for a few seconds and then stop it without looking at it. The final digit that measures time in $1/100^{th}$ of a second can be used as a random uniform digit from 0 to 9. A statistical analysis of 100 such random digits passed all of the standard diagnostic tests for randomness and uniformity (B. Inouye, personal communication).

knowledge of biology to separate plots or samples by some minimum distance or sampling interval to avoid dependence. However if we do not know all of the forces that can cause dependence, a random placement of plots beyond some minimum distance will ensure that the spacing of the plots is variable. Some plots will be relatively close, and some will be relatively far apart. Therefore, the effect of the dependence will be strong in some pairs of plots, weak in others, and non-existent in still others. Such variable effects may cancel one another and can reduce the chances that results are consistently biased by non-independence.

Finally, note that randomization and replication only are effective when they are used together. If we do not replicate, but simply assign randomly the control and treatment plots to the hillside or the valley, the design is still confounded (see Figure 6.6). Similarly, if we replicate the design, but assign all 10 of the controls to the valley and all 10 of the removals to the hillside, the design is also confounded (Figure 6.8). It is only when we use multiple plots and assign the treatments randomly that the confounding effect of temperature is removed from the design (see Figure 6.7). Indeed, it is fair to say that any unreplicated design is always going to be confounded with one or more environmental factors.[11]

Although the concept of randomization is straightforward, it must be applied at several stages in the design. First, randomization applies only to a well-defined, initially non-random sample space. The sample space doesn't simply mean the physical area from which replicates are sampled (although this is an important aspect of the sample space). Rather, the sample space refers to a set of elements that have experienced similar, though not identical, conditions.

Examples of a sample space might include individual cutthroat trout that are reproductively mature, lightfall gaps created by fires, old-fields abandoned 10–20

[11] Although confounding is easy to recognize in a field experiment of this sort, it may not be apparent that the same problem exists in laboratory and greenhouse experiments. If we rear insect larvae at high and low temperatures in two environmental chambers, this is a confounded design because all of the high temperature larvae are in one chamber and all of the low temperature larvae are in the other. If environmental factors other than temperature also differ between the chambers, their effects are confounded with temperature. The correct solution would be to rear each larva in its own separate chamber, thereby ensuring that each replicate is truly independent and that temperature is not confounded with other factors. But this sort of design simply is too expensive and wasteful of space ever to be used. Perhaps the argument can be made that environmental chambers and greenhouses really do differ only in temperature and no other factors, but that is only an assumption that should be tested explicitly. In many cases, the environment in environmental chambers is surprisingly heterogeneous, both within and between chambers. Potvin (2001) discusses how this variability can be measured and then used to design better laboratory experiments.

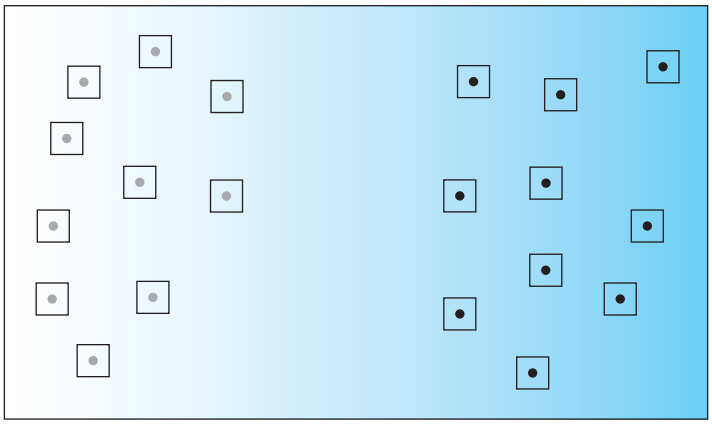

Figure 6.8 A replicated, but confounded, design. As in Figures 6.5, 6.6, and 6.7, the study establishes control and experimental nectar removal plots to evaluate the responses of foraging hummingbirds. Each square represents a replicate control plot (black dots) or nectar removal plot (gray dots). If treatments are replicated but not assigned randomly, the design still confounds treatments with underlying environmental gradients. Replication combined with randomization and sufficient spacing of replicates (see Figure 6.7) is the only safeguard against non-independence (see Figure 6.5) and confounding (see Figures 6.6 and 6.8).

years ago, or large bleached coral heads at 5–10 meters depth. Once this sample space has been defined clearly, sites, individuals, or replicates that meet the criteria should be chosen at random. As we noted in Chapter 1, the spatial and temporal boundaries of the study will dictate not only the sampling effort involved, but also the domain of inference for the conclusions of the study.

Once sites or samples are randomly selected, treatments should be assigned to them randomly, which ensures that different treatments are not clumped in space or confounded with environmental variables.[12] Samples should also be collected and treatments applied in a random sequence. That way, if environmental conditions change during the experiment, the results will not be

[12] If the sample size is too small, even a random assignment can lead to spatial clumping of treatments. One solution would be to set out the treatments in a repeated order (…123123…), which ensures that there is no clumping. However, if there is any non-independence among treatments, this design may exaggerate its effects, because Treatment 2 will always occur spatially between Treatments 1 and 3. A better solution would be to repeat the randomization and then statistically test the layout to ensure there is no clumping. See Hurlbert (1984) for a thorough discussion of the numerous hazards that can arise by failing to properly replicate and randomize ecological experiments.

confounded. For example, if you census all of your control plots first, and your field work is interrupted by a fierce thunderstorm, any impacts of the storm will be confounded with your manipulations because all of the treatment plots will be censused after the storm. These same provisos hold for non-experimental studies in which different plots or sites have to be censused. The caveat is that strictly random censusing in this way may be too inefficient because you will usually not be visiting neighboring sites in consecutive order. You may have to compromise between strict randomization and constraints imposed by sampling efficiency.

All methods of statistical analysis—whether they are parametric, Monte Carlo, or Bayesian (see Chapter 5)—rest on the assumption of random sampling at an appropriate spatial or temporal scale. You should get in the habit of using randomization whenever possible in your work.

Designing Effective Field Experiments and Sampling Studies

Here are some questions to ask when designing field experiments and sampling studies. Although some of these questions appear to be specific to manipulative experiments, they are also relevant to certain natural experiments, where "controls" might consist of plots lacking a particular species or set of abiotic conditions.

Are the Plots or Enclosures Large Enough to Ensure Realistic Results?

Field experiments that seek to control animal density must necessarily constrain the movement of animals. If the enclosures are too small, the movement, foraging, and mating behaviors of the animals may be so unrealistic that the results obtained will be uninterpretable or meaningless (MacNally 2000a). Try to use the largest plots or cages that are feasible and that are appropriate for the organism you are studying. The same considerations apply to sampling studies: the plots need to be large enough and sampled at an appropriate spatial scale to answer your question.

What Is the Grain and Extent of the Study?

Although much importance has been placed on the spatial scale of an experiment or a sampling study, there are actually two components of spatial scale that need to be addressed: grain and extent. **Grain** is the size of the smallest unit of study, which will usually be the size of an individual replicate or plot. **Extent** is the total area encompassed by all of the sampling units in the study. Grain and extent can be either large or small (Figure 6.9). There is no single combination of grain and extent that is necessarily correct. However, ecological studies with

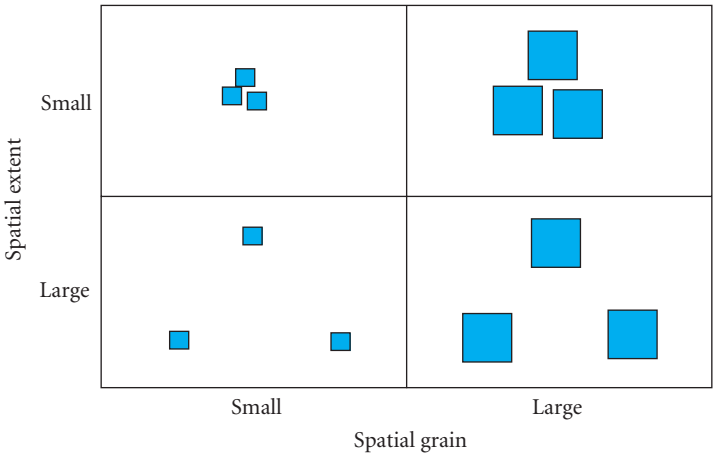

Figure 6.9 Spatial grain and spatial extent in ecological studies. Each square represents a single plot. Spatial grain measures the size of the sampling units, represented by small or large squares. Spatial extent measures the area encompassing all of the replicates of the study, represented by closely grouped or widely spaced squares.

both a small grain and a small extent, such as pitfall catches of beetles in a single forest plot, may sometimes be too limited in scope to allow for broad conclusions. On the other hand, studies with a large grain but a small extent, such as whole-lake manipulations in a single valley, may be very informative. Our own preference is for studies with a small grain, but a medium or large extent, such as ant and plant censuses in small plots (5 m × 5 m) across New England (Gotelli and Ellison 2002a,b) or eastern North America (Gotelli and Arnett 2000), or on small mangrove islands in the Caribbean (Farnsworth and Ellison 1996a). The small grain allows for experimental manipulations and observations taken at scales that are relevant to the organism, but the large extent expands the domain of inference for the results. In determining grain and extent, you should consider both the question you are trying to ask and the constraints on your sampling.

Does the Range of Treatments or Census Categories Bracket or Span the Range of Possible Environmental Conditions?

Many field experiments describe their manipulations as "bracketing or spanning the range of conditions encountered in the field." However, if you are trying to model climate change or altered environments, it may be necessary to also include conditions that are outside the range of those normally encountered in the field.

Have Appropriate Controls Been Established to Ensure that Results Reflect Variation Only in the Factor of Interest?

It is rare that a manipulation will change one and only one factor at a time. For example, if you surround plants with a cage to exclude herbivores, you have also altered the shading and moisture regime. If you simply compare these plants to unmanipulated controls, the herbivore effects are confounded with the differences in shading and moisture. The most common mistake in experimental designs is to establish a set of unmanipulated plots and then treat those as a control. Usually, an additional set of control plots that contain some minimal alteration will be necessary to properly control for the manipulations. In the example described above, an open-sided cage roof will allow herbivores access to plants, but will still include the shading effects of the cage. With this simple design of three treatments (Unmanipulated, Cage control, Herbivore exclusion), you can make the following contrasts:

1. *Unmanipulated versus Cage control.* This comparison reveals the extent to which shading and physical changes due to the cage per se are affecting plant growth and responses.
2. *Cage control versus Herbivore exclusion.* This comparison reveals the extent to which herbivory alters plant growth. Both the Control and Herbivore exclusion plots experience the shading effects of the cage, so any difference between them can be attributed to the effect of herbivores.
3. *Unmanipulated versus Herbivore exclusion.* This comparison measures the *combined effect* of both the herbivores and the shading on plant growth. Because the experiment is designed to measure only the herbivore effect, this particular comparison confounds treatment and caging effects.

In Chapter 10, we will explain how you use can use contrasts after analysis of variance to quantify these comparisons.

Have All Replicates Been Manipulated in the Same Way Except for the Intended Treatment Application?

Again, appropriate controls usually require more than lack of manipulation. If you have to push back plants to apply treatments, you should push back plants in the control plots as well (Salisbury 1963; Jaffe 1980). In a reciprocal transplant experiment with insect larvae, live animals may be sent via overnight courier to distant sites and established in new field populations. The appropriate control is a set of animals that are re-established in the populations from which they were collected. These animals will also have to receive the "UPS treatment" and be sent through the mail system to ensure they receive the same stress as the ani-

mals that were transplanted to distant sites. If you are not careful to ensure that all organisms are treated identically in your experiments, your treatments will be confounded with differences in handling effects (Cahill et al. 2000).

Have Appropriate Covariates Been Measured in Each Replicate?

Covariates are continuous variables (see Chapter 7) that potentially affect the response variable, but are not necessarily controlled or manipulated by the investigator. Examples include variation among plots in temperature, shade, pH, or herbivore density. Different statistical methods, such as analysis of covariance (see Chapter 10), can be used to quantify the effect of covariates.

However, you should avoid the temptation to measure every conceivable covariate in a plot just because you have the instrumentation (and the time) to do so. You will quickly end up with a dataset in which you have more variables measured than you have replicates, which causes additional problems in the analysis (Burnham and Anderson 2010). It is better to choose ahead of time the most biologically relevant covariates, measure only those covariates, and use sufficient replication. Remember also that the measurement of covariates is useful, but it is not a substitute for proper randomization and replication.

Summary

The sound design of an ecological experiment first requires a clear statement of the question being asked. Both manipulative and observational experiments can answer ecological questions, and each type of experiment has its own strengths and weaknesses. Investigators should consider the appropriateness of using a press versus a pulse experiment, and whether the replication will be in space (snapshot experiment), time (trajectory experiment), or both. Non-independence and confounding factors can compromise the statistical analysis of data from both manipulative and observational studies. Randomization, replication, and knowledge of the ecology and natural history of the organisms are the best safeguards against non-independence and confounding factors. Whenever possible, try to use at least 10 observations per treatment group. Field experiments usually require carefully designed controls to account for handling effects and other unintended alterations. Measurement of appropriate environmental covariates can be used to account for uncontrolled variation, although it is not a substitute for randomization and replication.