CHAPTER 7

A Bestiary of Experimental and Sampling Designs

In an experimental study, we have to decide on a set of biologically realistic manipulations that include appropriate controls. In an observational study, we have to decide which variables to measure that will best answer the question we have asked. These decisions are very important, and were the subject of Chapter 6. In this chapter we discuss specific designs for experimental and sampling studies in ecology and environmental science. The design of an experiment or observational study refers to how the replicates are physically arranged in space, and how those replicates are sampled through time. The design of the experiment is intimately linked to the details of replication, randomization, and independence (see Chapter 6). Certain kinds of designs have proven very powerful for the interpretation and analysis of field data. Other designs are more difficult to analyze and interpret. However, you cannot draw blood from a stone, and even the most sophisticated statistical analysis cannot rescue a poor design.

We first present a simple framework for classifying designs according to the types of independent and dependent variables. Next, we describe a small number of useful designs in each category. We discuss each design and the kinds of questions it can be used to address, illustrate it with a simple dataset, and describe the advantages and disadvantages of the design. The details of how to analyze data from these designs are postponed until Chapters 9–12.

The literature on experimental and sampling designs is vast (e.g., Cochran and Cox 1957; Winer 1991; Underwood 1997; Quinn and Keough 2002), and we present only a selective coverage in this chapter. We restrict ourselves to those designs that are practical and useful for ecologists and environmental scientists, and that have proven to be most successful in field studies.

Categorical versus Continuous Variables

We first distinguish between **categorical variables** and **continuous variables.** Categorical variables are classified into one of two or more unique categories. Ecological examples include sex (male, female), trophic status (producer, herbivore, carnivore), and habitat type (shade, sun). Continuous variables are measured on a continuous numerical scale; they can take on a range of real number or integer values. Examples include measurements of individual size, species richness, habitat coverage, and population density.

Many statistics texts make a further distinction between purely categorical variables, in which the categories are not ordered, and ranked (or ordinal) variables, in which the categories are ordered based on a numerical scale. An example of an ordinal variable would be a numeric score (0, 1, 2, 3, or 4) assigned to the amount of sunlight reaching the forest floor: 0 for 0–5% light; 1 for 6–25% light; 2 for 26–50% light; 3 for 51–75% light; and 4 for 76–100% light. In many cases, methods used for analyzing continuous data also can be applied to ordinal data. In a few cases, however, ordinal data are better analyzed with Monte Carlo methods, which were discussed in Chapter 5. In this book, we use the term *categorical variable* to refer to both ordered and unordered categorical variables.

The distinction between categorical and continuous variables is not always clear-cut; in many cases, the designation depends simply on how the investigator chooses to measure the variable. For example, a categorical habitat variable such as sun/shade could be measured on a continuous scale by using a light meter and recording light intensity in different places. Conversely, a continuous variable such as salinity could be classified into three levels (low, medium, and high) and treated as a categorical variable. Recognizing the kind of variable you are measuring is important because different designs are based on categorical and continuous variables.

In Chapter 2, we distinguished two kinds of random variables: discrete and continuous. What's the difference between discrete and continuous random variables on the one hand, and categorical and continuous variables on the other? Discrete and continuous random variables are mathematical functions for generating values associated with probability distributions. In contrast, categorical and continuous variables describe the kinds of data that we actually measure in the field or laboratory. Continuous variables usually can be modeled as continuous random variables, whereas both categorical and ordinal variables usually can be modeled as discrete random variables. For example, the categorical variable *sex* can be modeled as a binomial random variable; the numerical variable *height* can be modeled as normal random variable; and the ordinal variable *light reaching the forest floor* can be modeled as a binomial, Poisson, or uniform random variable.

Dependent and Independent Variables

After identifying the types of variables with which you are working, the next step is to designate **dependent** and **independent variables.** The assignment of dependent and independent variables implies an hypothesis of cause and effect that you are trying to test. The dependent variable is the **response variable** that you are measuring and for which you are trying to determine a cause or causes. In a scatterplot of two variables, the dependent or response variable is called the *Y* variable, and it usually is plotted on the **ordinate** (vertical or *y*-axis). The independent variable is the **predictor variable** that you hypothesize is responsible for the variation in the response variable. In the same scatterplot of two variables, the independent or predictor variable is called the *X* variable, and it usually is plotted on the **abscissa** (horizontal or *x*-axis).[1]

In an experimental study, you typically manipulate or directly control the levels of the independent variable and measure the response in the dependent variable. In an observational study, you depend on natural variation in the independent variable from one replicate to the next. In both natural and experimental studies, you don't know ahead of time the strength of the predictor variable. In fact, you are often testing the statistical null hypothesis that variation in the response variable is unrelated to variation in the predictor variable, and is no greater than that expected by chance or sampling error. The alternative hypothesis is that chance cannot entirely account for this variation, and that at least some of the variation can be attributed to the predictor variable. You also may be interested in estimating the size of the effect of the predictor or causal variable on the response variable.

Four Classes of Experimental Design

By combining variable types—categorical versus continuous, dependent versus independent—we obtain four different design classes (Table 7.1). When independent variables are continuous, the classes are either regression (continuous dependent variables) or logistic regression (categorical dependent variables). When independent variables are categorical, the classes are either ANOVA (continuous dependent variable) or tabular (categorical dependent variable). Not all designs fit nicely into these four categories. The analysis of covariance

[1] Of course, merely plotting a variable on the *x*-axis is does not guarantee that the it is actually the predictor variable. Particularly in natural experiments, the direction of cause and effect is not always clear, even though the measured variables may be highly correlated (see Chapter 6).

TABLE 7.1 **Four classes of experimental and sampling designs**

Dependent variable	Independent variable	
	Continuous	Categorical
Continuous	Regression	ANOVA
Categorical	Logistic regression	Tabular

Different kinds of designs are used depending on whether the independent and dependent variables are continuous or categorical. When both the dependent and the independent variables are continuous, a regression design is used. If the dependent variable is categorical and the independent variable is continuous, a logistic regression design is used. The analysis of regression designs is covered in Chapter 9. If the independent variable is categorical and the dependent variable is continuous, an analysis of variance (ANOVA) design is used. The analysis of ANOVA designs is described in Chapter 10. Finally, if both the dependent and independent variables are categorical, a tabular design is used. Analysis of tabular data is described in Chapter 11.

(ANCOVA) is used when there are two independent variables, one of which is categorical and one of which is continuous (the covariate). ANCOVA is discussed in Chapter 10. Table 7.1 categorizes univariate data, in which there is a single dependent variable. If, instead, we have a vector of correlated dependent variables, we rely on a multivariate analysis of variance (MANOVA) or other multivariate methods that are described in Chapter 12.

Regression Designs

When independent variables are measured on continuous numerical scales (see Figure 6.1 for an example), the sampling layout is a **regression design.** If the dependent variable is also measured on a continuous scale, we use linear or nonlinear regression models to analyze the data. If the dependent variable is measured on an ordinal scale (an ordered response), we use logistic regression to analyze the data. These three types of regression models are discussed in detail in Chapter 9.

SINGLE-FACTOR REGRESSION A regression design is simple and intuitive. Collect data on a set of independent replicates. For each replicate, measure both the predictor and the response variables. In an observational study, neither of the two variables is manipulated, and your sampling is dictated by the levels of natural variation in the independent variable. For example, suppose your hypothesis is that the density of desert rodents is controlled by the availability of seeds (Brown and Leiberman 1973). You could sample 20 independent plots, each

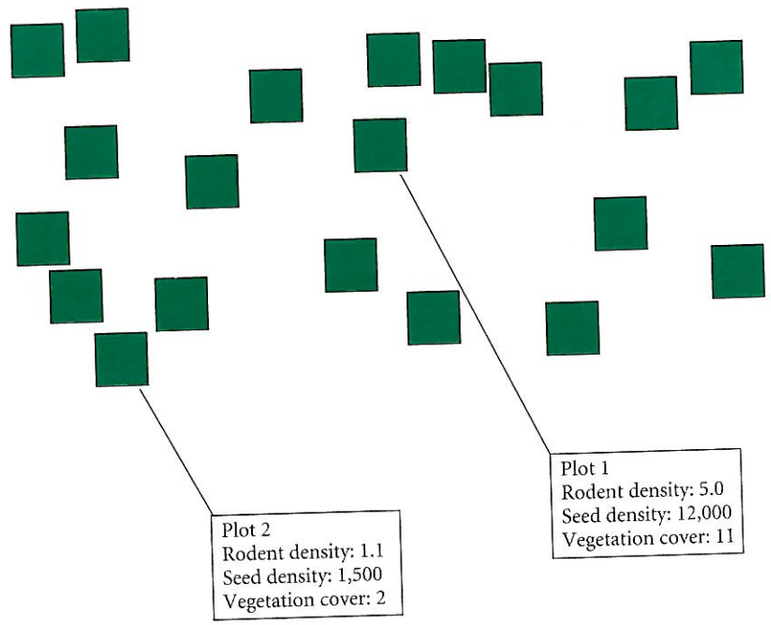
chosen to represent a different abundance level of seeds. In each plot, you measure the density of seeds and the density of desert rodents (Figure 7.1). The data are organized in a spreadsheet in which each row is a different plot, and each column is a different response or predictor variable. The entries in each row represent the measurements taken in a single plot.

In an experimental study, the levels of the predictor variable are controlled and manipulated directly, and you measure the response variable. Because your hypothesis is that seed density is responsible for desert rodent density (and not the other way around), you would manipulate seed density in an experimental study, either adding or removing seeds to alter their availability to rodents. In both the experimental study and the observational study, your assumption is that the predictor variable is a causal variable: changes in the value of the predictor (seed density) would *cause* a change in the value of the response (rodent density). This is very different from a study in which you would examine the correlation (statistical covariation) between the two variables. Correlation does not specify a cause-and-effect relationship between the two variables.[2]

In addition to the usual caveats about adequate replication and independence of the data (Chapter 6), two principles should be followed in designing a regression study:

1. *Ensure that the range of values sampled for the predictor variable is large enough to capture the full range to responses by the response variable.* If the predictor variable is sampled from too limited a range, there may appear to be a weak or nonexistent statistical relationship between the

[2] The sampling scheme needs to reflect the goals of the study. If the study is designed simply to document the relationship between seeds and rodent density, then a series of random plots can be selected, and **correlation** is used to explore the relationship between the two variables. However, if the hypothesis is that seed density is responsible for rodent density, then a series of plots that encompass a uniform range of seed densities should be sampled, and **regression** is used to explore the functional dependence of rodent abundance on seed density. Ideally, the sampled plots should differ from one another only in the density of seeds present. Another important distinction is that a true regression analysis assumes that the value of the independent variable is known exactly and is not subject to measurement error. Finally, standard linear regression minimizes residual deviations in the vertical (y) direction only, whereas correlation minimizes the perpendicular (x and y) distance of each point from the regression line (also referred to as Model II regression). The distinction between correlation and regression is subtle, and is often confusing because some statistics (such as the correlation coefficient) are identical for both kinds of analyses. See Chapter 9 for more details.



Plot number	Seeds/m^2	Vegetation cover (%)	Rodents/m^2
1	12,000	11	5.0
2	1,500	2	1.1
.	.	.	.
.	.	.	.
20	11,500	52	3.7

Figure 7.1 Spatial arrangement of replicates for a regression study. Each square represents a different 25-m^2 plot. Plots were sampled to ensure a uniform coverage of seed density (see Figures 7.2 and 7.3). Within each plot, the investigator measures rodent density (the response variable), and seed density and vegetation cover (the two predictor variables). The data are organized in a spreadsheet in which each row is a plot, and the columns are the measured variables within the plot.

predictor and response variables even though they are related (Figure 7.2). A limited sampling range makes the study susceptible to a Type II statistical error (failure to reject a false null hypothesis; see Chapter 4).

2. *Ensure that the distribution of predictor values is approximately uniform within the sampled range.* Beware of datasets in which one or two of the values of the predictor variable are very different in size from the others. These influential points can dominate the slope of the regression and

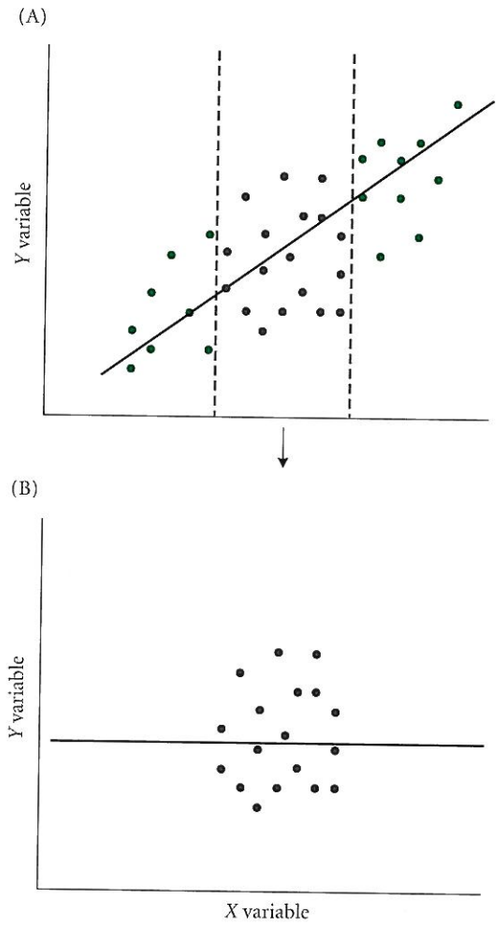


Figure 7.2 Inadequate sampling over a narrow range of the *X* variable can create a spuriously non-significant regression slope, even though *X* and *Y* are strongly correlated with one another. Each point represents a single replicate for which a value has been measured for both the *X* and the *Y* variables. Green circles represent possible data that were not collected for the analysis. Black circles represent the sample of replicates that were measured. (A) The full range of data. The solid line indicates the true linear relationship between the variables. (B) The regression line is fitted to the sample data. Because the *X* variable was sampled over a narrow range of values, there is limited variation in the resulting *Y* variable, and the slope of the fitted regression appears to be close to zero. Sampling over the entire range of the *X* variable will prevent this type of error.

generate a significant relationship where one does not really exist (Figure 7.3; see Chapter 8 for further discussion of such outliers). Sometimes influential data points can be corrected with a transformation of the predictor variable (see Chapter 8), but we re-emphasize that analysis cannot rescue a poor sampling design.

MULTIPLE REGRESSION The extension to multiple regression is straightforward. Two or more continuous predictor variables are measured for each replicate, along with the single response variable. Returning to the desert rodent example, you suspect that, in addition to seed availability, rodent density is also controlled by vegetation structure—in plots with sparse vegetation, desert rodents are vul-

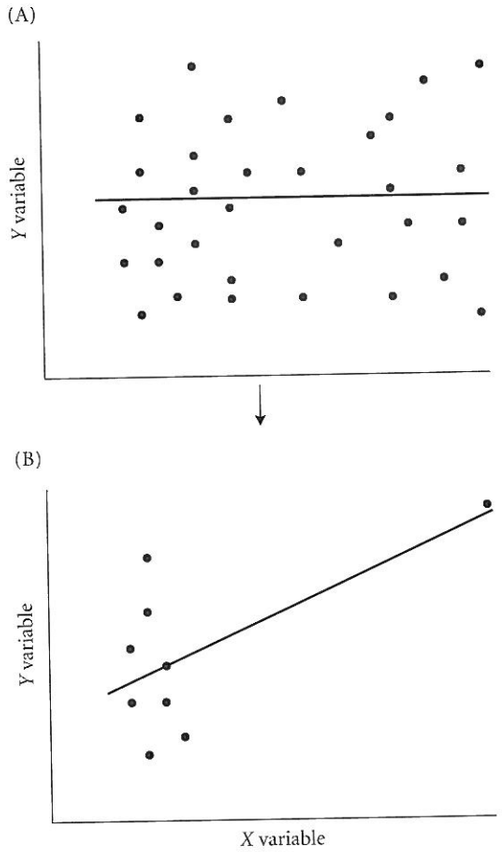


Figure 7.3 Failure to sample uniformly the entire range of a variable can lead to spurious results. As in Figure 7.2, each black point represents a single recorded observation; the green points represent unobserved X, Y pairs. (A) The solid line indicates the true linear relationship between the variables. This relationship would have been revealed if the X variable had been sampled uniformly. (B) The regression line is fitted to the sample data alone (i.e., just the black points). Because only a single datum with a large value of X was measured, this point has an inordinate influence on the fitted regression line. The result inaccurately suggests a positive relationship between the two variables.

nerable to avian predators (Abramsky et al. 1997). In this case, you would take three measurements in each plot: rodent density, seed density, and vegetation cover. Rodent density is still the response variable, and seed density and vegetation cover are the two predictor variables (Figure 7.1). Ideally, the different predictor variables should be independent of one another. As in simple regression designs, the different values of the predictor variables should be established evenly across the full range of possible values. This is straightforward in an experimental study, but rarely is achievable in an observational study. In an observational study, it is often the case that the predictor variables themselves will be correlated with each other. For example, plots with high vegetation density are likely to have high seed density. There may be few or no plots in which vegetation density is high and seed density is low (or vice versa). This **collinearity** makes it difficult to estimate accurately regression parameters[3] and to tease apart how much variation in the response variable is actually associated with each of the predictor variables.

As always, replication becomes important as we add more predictor variables to the analysis. Following the Rule of 10 (see Chapter 6), you should try to obtain at least 10 replicates for each predictor variable in your study. But in many studies, it is a lot easier to measure additional predictor variables than it is to obtain additional independent replicates. However, you should avoid the temptation to measure everything that you can just because it is possible. Try to select variables that are biologically important and relevant to the hypothesis or question you are asking. It is a mistake to think that a model selection algorithm, such as stepwise multiple regression, can identify reliably the "correct" set of predictor variables from a large dataset (Burnham and Anderson 2002). Moreover, large datasets often suffer from **multicollinearity**: many of the predictor variables are correlated with one another (Graham 2003).

ANOVA Designs

If your predictor variables are categorical (ordered or unordered) and your response variables are continuous, your design is called an **ANOVA** (for **AN**alysis **O**f **VA**riance). ANOVA also refers to the statistical analysis of these types of designs (see Chapter 10).

TERMINOLOGY ANOVA is rife with terminology. **Treatments** refer to the different categories of the predictor variables that are used. In an experimental study, the treatments represent the different manipulations that have been performed. In an observational study, the treatments represent the different groups that are being compared. The number of treatments in a study equals the number of categories being compared. Within each treatment, multiple observations will be made, and each of these observations is a **replicate**. In standard ANOVA designs, each replicate should be independent, both statistically and biologically, of the other replicates within and among treatments. Later in this

[3] In fact, if one of the predictor variables can be described as a perfect linear function of the other one, it is not even algebraically possible to solve for the regression coefficients. Even when the problem is not this severe, correlations among predictor variables make it difficult to test and compare models. See MacNally (2000b) for a discussion of correlated variables and model-building in conservation biology.

chapter, we will discuss certain ANOVA designs that relax the assumption of independence among replicates.

We also distinguish between **single-factor designs** and **multifactor designs**. In a single-factor design, each of the treatments represents variation in a single predictor variable or **factor**. Each value of the factor that represents a particular treatment is called a **treatment level**. For example, a single-factor ANOVA design could be used to compare growth responses of plants raised at 4 different levels of nitrogen, or the growth responses of 5 different plant species to a single level of nitrogen. The treatment groups may be ordered (e.g., 4 nitrogen levels) or unordered (e.g., 5 plant species).

In a multifactor design, the treatments cover two (or more) different factors, and each factor is applied in combination in different treatments. In a multifactor design, there are different levels of the treatment for each factor. As in the single-factor design, the treatments within each factor may be either ordered or unordered. For example, a two-factor ANOVA design would be necessary if you wanted to compare the responses of plants to 4 levels of nitrogen (Factor 1) *and* 4 levels of phosphorus (Factor 2). In this design, each of the $4 \times 4 = 16$ treatment levels represents a different combination of nitrogen level and phosphorus level. Each combination of nutrients is applied to all of the replicates within the treatment (Figure 7.4).

Although we will return to this topic later, it is worth asking at this point what the advantage is of using a two-factor design. Why not just run two separate experiments? For example, you could test the effects of phosphorus in a one-way ANOVA design with 4 treatment levels, and you could test the effects of nitrogen in a separate one-way ANOVA design, also with 4 treatment levels. What is the advantage of using a two-way design with 16 phosphorus–nitrogen treatment combinations in a single experiment?

One advantage of the two-way design is efficiency. It is likely to be more cost-effective to run a single experiment—even one with 16 treatments—than to run two separate experiments with 4 treatments each. A more important advantage is that the two-way design allows you to test both for main effects (e.g., the effects of nitrogen and phosphorus on plant growth) and for interaction effects (e.g., interactions between nitrogen and phosphorus).

The **main effects** are the additive effects of each level of one treatment averaged over all of the levels of the other treatment. For example, the additive effect of nitrogen would represent the response of plants at each nitrogen level, averaged over the responses to the phosphorus levels. Conversely, the additive effect of phosphorus would be measured as the response of plants at each phosphorus level, averaged over the responses to the different nitrogen levels.

Nitrogen treatment (*one-way layout*)			
0.00 mg	0.10 mg	0.50 mg	1.00 mg
10	10	10	10

Phosphorous treatment (*one-way layout*)			
0.00 mg	0.05 mg	0.10 mg	0.25 mg
10	10	10	10

(*Simultaneous N and P treatments in a two-way layout*)

		Nitrogen treatment			
		0.00 mg	0.10 mg	0.50 mg	1.00 mg
Phosphorus treatment	0.00 mg	10	10	10	10
	0.05 mg	10	10	10	10
	0.10 mg	10	10	10	10
	0.25 mg	10	10	10	10

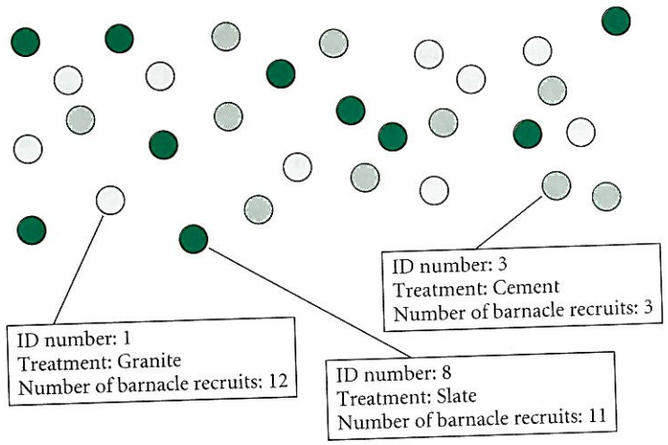
Figure 7.4 Treatment combinations in single-factor designs (upper two panels) and in a two-factor design (lower panel). In all designs, the number in each cell indicates the number of independent replicate plots to be established. In the two single-factor designs (one-way layouts), the four treatment levels represent four different nitrogen or phosphorous concentrations (mg/L). The total sample size is 40 plots in each single-factor experiment. In the two-factor design, the $4 \times 4 = 16$ treatments represent different combinations of nitrogen *and* phosphorous concentrations that are applied simultaneously to a replicate plot. This fully crossed two-factor ANOVA design with 10 replicates per treatment combination would require a total sample size of 160 plots. See Figures 7.9 and 7.10 for other examples of a crossed two-factor design.

Interaction effects represent unique responses to particular treatment combinations that cannot be predicted simply from knowing the main effects. For example, the growth of plants in the high nitrogen–high phosphorus treatment might be synergistically greater than you would predict from knowing the simple additive effects of nitrogen and phosphorus at high levels. Interaction effects are frequently the most important reason for using a factorial design. Strong interactions are the driving force behind much ecological and evolutionary change, and often are more important than the main effects. Chapter 10 will discuss analytical methods and interaction terms in more detail.

SINGLE-FACTOR ANOVA The single-factor ANOVA is one of the simplest, but most powerful, experimental designs. After describing the basic one-way layout, we also explain the randomized block and nested ANOVA designs. Strictly speaking, the randomized block and nested ANOVA are two-factor designs, but the second factor (blocks, or subsamples) is included only to control for sampling variation and is not of primary interest.

The **one-way layout** is used to compare means among two or more treatments or groups. For example, suppose you want to determine whether the recruitment of barnacles in the intertidal zone of a shoreline is affected by different kinds of rock substrates (Caffey 1982). You could start by obtaining a set of slate, granite, and concrete tiles. The tiles should be identical in size and shape, and differ only in material. Following the Rule of 10 (Chapter 6), set out 10 replicates of each substrate type ($N = 30$ total). Each replicate is placed in the mid-intertidal zone at a set of spatial coordinates that were chosen with a random number generator (Figure 7.5).

After setting up the experiment, you return 10 days later and count the number of new barnacle recruits inside a 10 cm × 10 cm square centered in the middle of each tile. The data are organized in a spreadsheet in which each row is a



ID number	Treatment	Replicate	Number of barnacle recruits
1	Granite	1	12
2	Slate	1	10
3	Cement	1	3
4	Granite	2	14
5	Slate	2	10
6	Cement	2	8
7	Granite	3	11
8	Slate	3	11
9	Cement	3	7
.	.	.	.
.	.	.	.
30	Cement	10	8

Figure 7.5 Example of a one-way layout. This experiment is designed to test for the effect of substrate type on barnacle recruitment in the rocky intertidal (Caffey 1982). Each circle represents an independent rock substrate. There are 10 randomly placed replicates of each of three treatments, represented by the three shades of green. The number of barnacle recruits is sampled from a 10-cm square in the center of each rock surface. The data are organized in a spreadsheet in which each row is an independent replicate. The columns indicate the ID number of each replicate (1–30), the treatment group (Cement, Slate, or Granite), the replicate number within each treatment (1–10), and the number of barnacle recruits (the response variable).

replicate. The first few columns contain identifying information associated with the replicate, and the last column of the spreadsheet gives the number of barnacles that recruited into the square. Although the details are different, this is the same layout used in the study of ant density described in Chapter 5: multiple, independent replicate observations are obtained for each treatment or sampling group.

The one-way layout is one of the simplest but most powerful experimental designs, and it can readily accommodate studies in which the number of replicates per treatment is not identical (unequal sample sizes). The one-way layout allows you to test for differences among treatments, as well as to test more specific hypotheses about which particular treatment group means are different and which are similar (see "Comparing Means" in Chapter 10).

The major disadvantage of the one-way layout is that it does not explicitly accommodate environmental heterogeneity. Complete randomization of the replicates within each treatment implies that they will sample the entire array of background conditions, all of which may affect the response variable. On the one hand, this is a good thing because it means that the results of the experiment can be generalized across all of these environments. On the other hand, if the environmental "noise" is much stronger than the "signal" of the treatment, the experiment will have low power; the analysis may not reveal treatment differences unless there are many replicates. Other designs, including the randomized block and the two-way layout, can be used to accommodate environmental variability.

A second, more subtle, disadvantage of the one-way layout is that it organizes the treatment groups along a single factor. If the treatments represent distinctly different kinds of factors, then a two-way layout should be used to tease apart main effects and interaction terms. Interaction terms are especially important because the effect of one factor often depends on the levels of another. For example, the pattern of recruitment onto different substrates may depend on the levels of a second factor (such as predator density).

RANDOMIZED BLOCK DESIGNS One effective way to incorporate environmental heterogeneity is to modify the one-way ANOVA and use a **randomized block design**. A **block** is a delineated area or time period within which the environmental conditions are relatively homogeneous. Blocks may be placed randomly or systematically in the study area, but they should be arranged so that environmental conditions are more similar within blocks than between them.

Once the blocks are established, replicates will still be assigned randomly to treatments, but there is a restriction on the randomization: a single replicate from each of the treatments is assigned to each block. Thus, in a simple ran-

domized block design, each block contains exactly one replicate of all the treatments in the experiment. Within each block, the placement of the treatment replicates should be randomized. Figure 7.6 illustrates the barnacle experiment laid out as a randomized block design. Because there are 10 replicates, there are

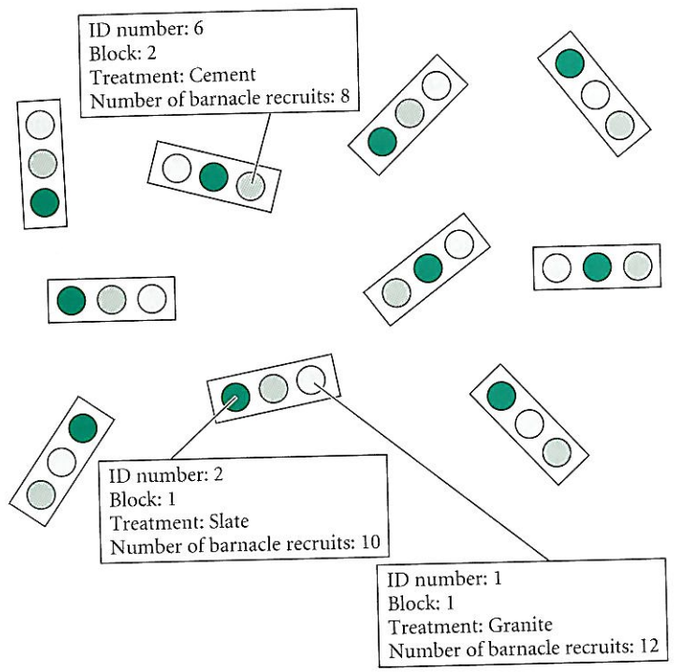


Figure 7.6 Example of a randomized block design. The 10 replicates of each of the three treatments are grouped in blocks—physical groupings of one replicate of each of the three treatments. Both the placement of blocks and placement of treatments within blocks are randomized. Data organization in the spreadsheet is identical to the one-way layout (Figure 7.5), but the replicate column is replaced by a column indicating the block with which each replicate is associated.

ID number	Treatment	Block	Number of barnacle recruits
1	Granite	1	12
2	Slate	1	10
3	Cement	1	3
4	Granite	2	14
5	Slate	2	10
6	Cement	2	8
7	Granite	3	11
8	Slate	3	11
9	Cement	3	7
.	.	.	.
.	.	.	.
30	Cement	10	8

10 blocks (fewer, if you replicate within each block), and each block will contain one replicate of each of the three treatments. The spreadsheet layout for these data is the same as for the one-way layout, except the replicate column is now replaced by a column indicating the block.

Each block should be small enough to encompass a relatively homogenous set of conditions. However, each block must also be large enough to accommodate a single replicate of each of the treatments. Moreover, there must be room within the block to allow sufficient spacing between replicates to ensure their independence (see Figure 6.5). The blocks themselves also have to be far enough apart from one another to ensure independence of replicates among blocks.

If there are geographic gradients in environmental conditions, then each block should encompass a small interval of the gradient. For example, there are strong environmental gradients along a mountainside, so we might set up an experiment with three blocks, one each at high, medium, and low elevation (Figure 7.7A). But it would not be appropriate to create three blocks that run "across the grain" from high to low elevation (Figure 7.7B); each block encompasses conditions that are too heterogeneous. In other cases, the environmental variation may be patchy, and the blocks should be arranged to reflect that patchiness. For example, if an experiment is being conducted in a wetland complex, each semi-isolated fen could be treated as a block. Finally, if the spatial organization of environmental heterogeneity is unknown, the blocks can be arranged randomly within the study area.[4]

The randomized block design is an efficient and very flexible design that provides a simple control for environmental heterogeneity. It can be used to control for environmental gradients and patchy habitats. As we will see in Chapter 10, when environmental heterogeneity is present, the randomized block design is more efficient than a completely randomized one-way layout, which may require a great deal more replication to achieve the same statistical power.

[4] The randomized block design allows you to set up your blocks to encompass environmental gradients in a single spatial dimension. But what if the variation occurs in two dimensions? For example, suppose there is a north-to-south moisture gradient in a field, but also an east-to-west gradient in predator density. In such cases, more complex randomized block designs can be used. For example, the **Latin square** is a block design in which the n treatments are placed in the field in an $n \times n$ square; each treatment appears exactly once in every row and once in every column of the layout. Sir Ronald Fisher (see Footnote 5, Chapter 5) pioneered these kinds of designs for agricultural studies in which a single field is partitioned and treatments applied to the contiguous subplots. These designs have not been used often by ecologists because the restrictions on randomization and layout are difficult to achieve in field experiments.

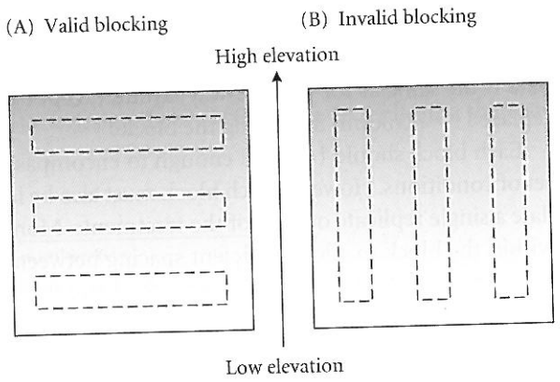


Figure 7.7 Valid and invalid blocking designs. (A) Three properly oriented blocks, each encompassing a single elevation on a mountainside or other environmental gradient. Environmental conditions are more similar within than among blocks. (B) These blocks are oriented improperly, going "across the grain" of the elevational gradient. Conditions are as heterogeneous within the blocks as between them, so no advantage is gained by blocking.

The randomized block design is also useful when your replication is constrained by space or time. For example, suppose you are running a laboratory experiment on algal growth with 8 treatments and you want to complete 10 replicates per treatment. However, you have enough space in your laboratory to run only 12 replicates at a time. What can you do? You should run the experiment in blocks, in which you set up a single replicate of each of the 8 treatments. After the result is recorded, you set up the experiment again (including another set of randomizations for treatment establishment and placement) and continue until you have accumulated 10 blocks. This design controls for inevitable changes in environmental conditions that occur in your laboratory through time, but still allows for appropriate comparison of treatments. In other cases, the limitation may not be space, but organisms. For example, in a study of mating behavior of fish, you may have to wait until you have a certain number of sexually mature fish before you can set up and run a single block of the experiment. In both examples, the randomized block design is the best safeguard against variation in background conditions during the course of your experiment.

Finally, the randomized block design can be adapted for a **matched pairs layout.** Each block consists of a group of individual organisms or plots that have been deliberately chosen to be most similar in background characteristics. Each replicate in the group receives one of the assigned treatments. For example, in a simple experimental study of the effects of abrasion on coral growth, a pair of coral heads of similar size would be considered a single block. One of the coral heads would be randomly assigned to the control group, and the other would be assigned to the abrasion group. Other matched pairs would be chosen in the same way and the treatments applied. Even though the individuals in each pair are not part of a spatial or a temporal block, they are probably going to be more similar than individuals in other such blocks because they have been matched on the basis of colony size or other characteristics. For this reason, the analysis will use a randomized block design. The matched pairs approach is a very effective method when the responses of the replicates potentially are very heterogeneous. Matching the individuals controls for that heterogeneity, making it easier to detect treatment effects.

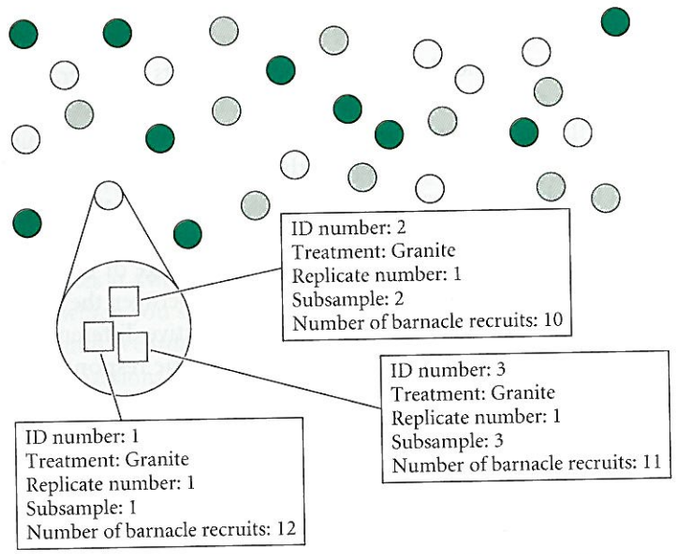
There are four disadvantages to the randomized block design. The first is that there is a statistical cost to running the experiment with blocks. If the sample size is small and the block effect is weak, the randomized block design is less powerful than a simple one-way layout (see Chapter 10). The second disadvantage is that if the blocks are too small you may introduce non-independence by physically crowding the treatments together. As we discussed in Chapter 6, randomizing the placement of the treatments within the block will help with this problem, but won't eliminate it entirely. The third disadvantage of the randomized block design is that if any of the replicates are lost, the data from that block cannot be used unless the missing values can be estimated indirectly.

The fourth—and most serious—disadvantage of the randomized block design is that it assumes there is no interaction between the blocks and the treatments. The blocking design accounts for additive differences in the response variable and assumes that the rank order of the responses to the treatment does not change from one block to the next. Returning to the barnacle example, the randomized block model assumes that if recruitment in one of the blocks is high, all of the observations in that block will have elevated recruitment. However, the treatment effects are assumed to be consistent from one block to the next, so that the rank order of barnacle recruitment among treatments (Granite > Slate > Cement) is the same, regardless of any differences in the overall recruitment levels among blocks. But suppose that in some blocks recruitment is highest on the cement substrate and in other blocks it is highest on the granite substrate. In this case, the randomized block design may fail to properly characterize the main treatment effects. For this reason, some authors (Mead 1988, Underwood 1997) have argued that the simple randomized block design should not be used unless there is replication within blocks. With replication, the design becomes a two-factor analysis of variance, which we discuss below.

Replication within blocks will indeed tease apart main effects, block effects, and the interaction between blocks and treatments. Replication will also address the problem of missing or lost data from within a block. However, ecologists often do not have the luxury of replication within blocks, particularly when the blocking factor is not of primary interest. The simple randomized block

design (without replication) will at least capture the additive component (often the most important component) of environmental variation that would otherwise be lumped with pure error in a simple one-way layout.

NESTED DESIGNS A **nested design** refers to any design in which there is subsampling within each of the replicates. We will illustrate it with the barnacle example. Suppose that, instead of measuring recruitment for a replicate in a single 10 cm × 10 cm square, you decided to take three such measurements for each of the 30 tiles in the study (Figure 7.8). Although the number of replicates has



ID number	Treatment	Replicate number	Subsample	Number of barnacle recruits
1	Granite	1	1	12
2	Granite	1	2	10
3	Granite	1	3	11
4	Slate	2	1	14
5	Slate	2	2	10
6	Slate	2	3	7
7	Cement	3	1	5
8	Cement	3	2	6
9	Cement	3	3	10
.	.	.	.	.
.	.	.	.	.
90	Cement	30	3	6

Figure 7.8 Example of a nested design. The study is the same as that shown in Figures 7.5 and 7.6. The layout is identical to the one-way layout in Figure 7.5, but here three subsamples are taken for each independent replicate. In the spreadsheet, an additional column is added to indicate the subsample number, and the total number of observations is increased from 30 to 90.

not increased, the number of observations has increased from 30 to 90. In the spreadsheet for these data, each row now represents a different subsample, and the columns indicate from which replicate and from which treatment the subsample was taken.

This is the first design in which we have included subsamples that are clearly not independent of one another. What is the rationale for such a sampling scheme? The main reason is to increase the precision with which we estimate the response for each replicate. Because of the Law of Large Numbers (Chapter 3), the more subsamples we use, the more precisely we will estimate the mean for each replicate. The increase in precision should increase the power of the test.

There are three advantages to using a nested design. The first advantage, as we noted, is that subsampling increases the precision of the estimate for each replicate in the design. Second, the nested design allows you to test two hypotheses: first, is there variation among treatments? And, second, is there variation among the replicates *within* a treatment? The first hypothesis is equivalent to a one-way design that uses the *subsample averages* as the observation for each replicate. The second hypothesis is equivalent to a one-way design that uses the subsamples to test for differences among replicates *within* treatments.[5]

Finally, the nested design can be extended to a hierarchical sampling design. For example, you could, in a single study, census subsamples nested within replicates, replicates nested within intertidal zones, intertidal zones nested within shores, shores nested within regions, and even regions nested within continents (Caffey 1985). The reason for carrying out this kind of sampling is that the variation in the data can be partitioned into components that represent each of the hierarchical levels of the study (see Chapter 10). For example, you might be able to show that 80% of the variation in the data occurs at the level of intertidal zones within shores, but only 2% can be attributed to variation among shores within a region. This would mean that barnacle density varies strongly from the high to the low intertidal, but doesn't vary much from one shoreline to the next. Such statements are useful for assessing the relative importance of different mechanisms in producing pattern (Petraitis 1998; see also Figure 4.6).

Nested designs potentially are dangerous in that they are often analyzed incorrectly. One of the most serious and common mistakes in ANOVA is for

[5] You can think of this second hypothesis as a one-way design at a lower hierarchical level. For example, suppose you used the data only from the four replicates of the granite treatment. Consider each replicate as a different "treatment" and each subsample as a different "replicate" for that treatment. The design is now a one-way design that compares the replicates of the granite treatment.

investigators to treat each subsample as an independent replicate and analyze the nested design as a one-way design (Hurlbert 1984). The non-independence of the subsamples artificially boosts the sample size (by threefold in our example, in which we took three subsamples from each tile) and badly inflates the probability of a Type I statistical error (i.e., falsely rejecting a true null hypothesis). A second, less serious problem, is that the nested design can be difficult or even impossible to analyze properly if the sample sizes are not equal in each group. Even with equal numbers of samples and subsamples, nested sampling in more complex layouts, such as the two-way layout or the split-plot design, can be tricky to analyze; the simple default settings for statistical software usually are not appropriate.

But the most serious disadvantage of the nested design is that it often represents a case of misplaced sampling effort. As we will see in Chapter 10, the power of ANOVA designs depends much more on the number of independent replicates than on the precision with which each replicate is measured. It is a much better strategy to invest your sampling effort in obtaining more independent replicates than subsampling within each replicate. By carefully specifying your sampling protocol (e.g., "only undamaged fruits from uncrowded plants growing in full shade"), you may be able to increase the precision of your estimates more effectively than by repeated subsampling.

That being said, you should certainly go ahead and subsample if it is quick and cheap to do so. However, our advice is that you then average (or pool) those subsamples so that you have a single observation for each replicate and then treat the experiment as a one-way design. As long as the numbers aren't too unbalanced, averaging also can alleviate problems of unequal sample size among subsamples and improve the fit of the errors to a normal distribution. It is possible, however, that after averaging among subsamples within replicates you no longer have sufficient replicates for a full analysis. In that case, you need a design with more replicates that are truly independent. Subsampling is no solution to inadequate replication!

MULTIPLE-FACTOR DESIGNS: TWO-WAY LAYOUT Multifactor designs extend the principles of the one-way layout to two or more treatment factors. Issues of randomization, layout, and sampling are identical to those discussed for the one-way, randomized block, and nested designs. Indeed, the only real difference in the design is in the assignment of the treatments to two or more factors instead of to a single factor. As before, the factors can represent either ordered or unordered treatments.

Returning again to the barnacle example, suppose that, in addition to substrate effects, you wanted to test the effects of predatory snails on barnacle recruitment. You could set up a second one-way experiment in which you established

four treatments: unmanipulated, cage control,[6] predator exclusion, and predator inclusion. Instead of running two separate experiments, however, you decide to examine both factors in a single experiment. Not only is this a more efficient use of your field time, but also the effect of predators on barnacle recruitment might differ depending on the substrate type. Therefore, you establish treatments in which you simultaneously apply a different substrate and a different predation treatment.

This is an example of a **factorial design** in which two or more factors are tested simultaneously in one experiment. The key element of a proper factorial design is that the treatments are **fully crossed** or **orthogonal**: every treatment level of the first factor (substrate) must be represented with every treatment level of the second factor (predation; Figure 7.9). Thus, the two-factor experiment has $3 \times 4 = 12$ distinct treatment combinations, as opposed to only 3 treatments for the single-factor substrate experiment or 4 treatments for the single-factor predation experiment. Notice that each of these single-factor experiments would be restricted to only one of the treatment combinations of the other factor. In other words, the substrate experiment that we described above was conducted with the unmanipulated predation treatment, and the predation treatment would be conducted on only a single substrate type. Once we have determined the treatment combinations, the physical set up of the experiment would be the same as for a one-way layout with 12 treatment combinations (Figure 7.10).

In the two-factor experiment, it is critical that all of the crossed treatment combinations be represented in the design. If some of the treatment combinations are missing, we end up with a confounded design. As an extreme example, suppose we set up only the granite substrate–predator exclusion treatment and the slate substrate–predator inclusion treatment. Now the predator effect is confounded with the substrate effect. Whether the results are statistically significant or not, we cannot tease apart whether the pattern is due to the effect of the predator, the effect of the substrate, or the interaction between them.

This example highlights an important difference between manipulative experiments and observational studies. In the observational study, we would gather data on variation in predator and prey abundance from a range of samples.



Cage and cage control

[6] In a cage control, investigators attempt to mimic the physical conditions generated by the cage, but still allow organisms to move freely in and out of the plot. For example, a cage control might consist of a mesh roof (placed over a plot) that allows predatory snails to enter from the sides. In an exclusion treatment, all predators are removed from a mesh cage, and in the inclusion treatment, predators are placed inside each mesh cage. The accompanying figure illustrates a cage (upper panel) and cage control (lower panel) in a fish exclusion experiment in a Venezuelan stream (Flecker 1996).

Substrate treatment (*one-way layout*)		
Granite	Slate	Cement
10	10	10

Predator treatment (*one-way layout*)			
Unmanipulated	Control	Predator exclusion	Predator inclusion
10	10	10	10

(*Simultaneous predator and substrate treatments in a two-way layout*)		Substrate treatment		
		Granite	Slate	Cement
Predator treatment	Unmanipulated	10	10	10
	Control	10	10	10
	Predator exclusion	10	10	10
	Predator inclusion	10	10	10

Figure 7.9 Treatment combinations in two single-factor designs and in a fully crossed two-factor design. This experiment is designed to test for the effect of substrate type (Granite, Slate, or Cement) and predation (Unmanipulated, Control, Predator exclusion, Predator inclusion) on barnacle recruitment in the rocky intertidal. The number 10 indicates the total number of replicates in each treatment. The three shades of the circles represent the three substrate treatments, and the patterns of the squares represent the four predation treatments. The two upper panels illustrate two one-way designs, in which only one of the two factors is systematically varied. In the two-factor design (lower panel), the $4 \times 3 = 12$ treatments represent different combinations of substrate *and* predation. The symbol in each cell indicates the combination of predation and substrate treatment that is applied.

But predators often are restricted to only certain microhabitats or substrate types, so that the presence or absence of the predator is indeed naturally confounded with differences in substrate type. This makes it difficult to tease apart cause and effect (see Chapter 6). The strength of multifactor field experiments is that they break apart this natural covariation and reveal the effects of multiple factors separately *and* in concert. The fact that some of these treatment combinations may be artificial and rarely, if ever, found in nature, actually is a strength of the experiment: it reveals the independent contribution of each factor to the observed patterns.

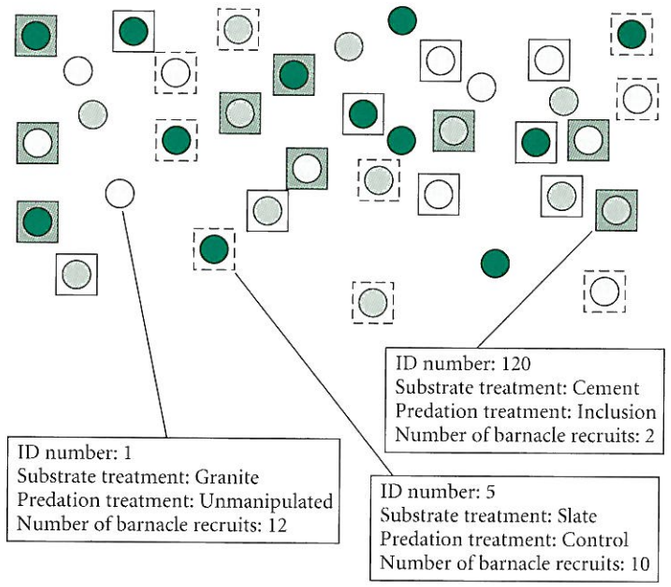


Figure 7.10 Example of a two-way design. Treatment symbols are given in Figure 7.9. The spreadsheet contains columns to indicate which substrate treatment and which predation treatment were applied to each replicate. The entire design includes $4 \times 3 \times 10 = 120$ replicates total, but only 36 replicates (three per treatment combination) are illustrated.

ID number	Substrate treatment	Predation treatment	Number of barnacle recruits
1	Granite	Unmanipulated	12
2	Slate	Unmanipulated	10
3	Cement	Unmanipulated	8
4	Granite	Control	14
5	Slate	Control	10
6	Cement	Control	8
7	Granite	Predator exclusion	50
8	Slate	Predator exclusion	68
9	Cement	Predator exclusion	39
.	.	.	.
.	.	.	.
120	Cement	Predator inclusion	2

The key advantage of two-way designs is the ability to tease apart main effects and interactions between two factors. As we will discuss in Chapter 10, the interaction term represents the non-additive component of the response. The interaction measures the extent to which different treatment combinations act additively, synergistically, or antagonistically.

Perhaps the main disadvantage of the two-way design is that the number of treatment combinations can quickly become too large for adequate replica-

tion. In the barnacle predation example, 120 total replicates are required to replicate each treatment combination 10 times.

As with the one-way layout, a simple two-way layout does not account for spatial heterogeneity. This can be handled by a simple randomized block design, in which each block contains exactly one of the treatment combinations. Alternatively, if you replicate all of the treatments within each block, this becomes a three-way design, with the blocks forming the third factor in the analysis.

A final limitation of two-way designs is that it may not be possible to establish all orthogonal treatment combinations. It is somewhat surprising that for many common ecological experiments, the full set of treatment combinations may not be feasible or logical. For example, suppose you are studying the effects of competition between two species of salamanders on salamander survival rate. You decide to use a simple two-way design in which each species represents one of the factors. Within each factor, the two treatments are the presence or absence of the species. This fully crossed design yields four treatments (Table 7.2). But what are you going to measure in the treatment combination that has neither Species A nor Species B? By definition, there is nothing to measure in this treatment combination. Instead, you will have to establish the other three treatments ([Species A Present, Species B Absent], [Species A Absent, Species B Present], [Species A Present, Species B Present]) and analyze the design as a one-way ANOVA. The two-way design is possible only if we change the response variable. If the response variable is the abundance of prey that are eaten by salamanders, rather than salamander survivorship, we can then establish the treatment with no salamanders of either species and measure prey levels in the fully crossed two-way layout. Of course, this experiment now asks an entirely different question.

TABLE 7.2 Treatment combinations in a two-way layout for simple species addition and removal experiments

Species B	Species A	
	Absent	Present
Absent	10	10
Present	10	10

The entry in each cell is the number of replicates of each treatment combination. If the response variable is some property of the species themselves (e.g., survivorship, growth rate), then the treatment combination Species A Absent–Species B Absent (boxed) is not logically possible, and the analysis will have to use a one-way layout with three treatment groups (Species A Present–Species B Present, Species A Present–Species B Absent, and Species A Absent–Species B Present). If the response variable is some property of the environment that is potentially affected by the species (e.g., prey abundance, pH), then all four treatment combinations can be used and analyzed as a two-way ANOVA with two orthogonal factors (Species A and Species B), each with two treatment levels (Absent, Present).

Two-species competition experiments like our salamander example have a long history in ecological and environmental research (Goldberg and Scheiner 2001). A number of subtle problems arise in the design and analysis of two-species competition experiments. These experiments attempt to distinguish between a focal species, for which the response variable is measured, an associative species, whose density is manipulated, and background species, which may be present, but are not experimentally manipulated.

The first issue is what kind of design to use: **additive, substitutive**, or **response surface** (Figure 7.11; Silvertown 1987). In an additive design, the density of the focal species is kept constant while the density of the experimental species is varied. However, this design confounds both density and frequency effects. For example, if we compare a control plot (5 individuals of Species A, 0 individuals of Species B) to an addition plot (5 individuals of Species A, 5 individuals of Species B), we have confounded total density (10 individuals) with the presence of the competitor (Underwood 1986, Bernardo et al. 1995). On the other hand, some

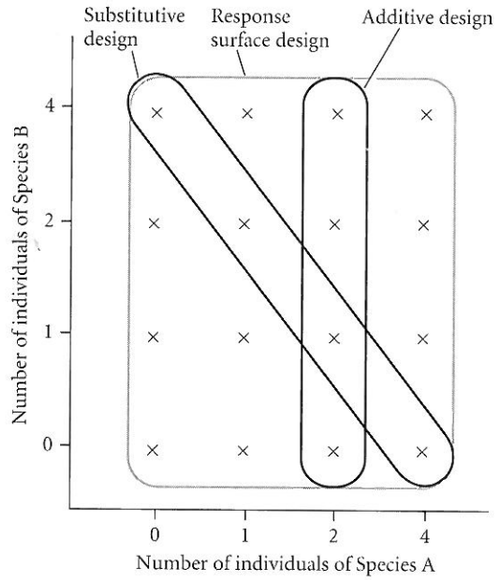


Figure 7.11 Experimental designs for competition experiments. The abundance of Species A and B are each set at 0, 1, 2, or 4 individuals. Each × indicates a different treatment combination. In an additive design, the abundance of one species is fixed (2 individuals of Species A) and the abundance of the competitor is varied (0, 1, 2, or 4 individuals of Species B). In a substitutive design, the total abundance of both competitors is held constant at 4 individuals, but the species composition in the different treatments is altered (0,4; 1,3; 2,2; 3,1; 4,0). In a response surface design, all abundance combinations of the two competitors are established in different treatments ($4 \times 4 = 16$ treatments). The response surface design is preferred because it follows the principle of a good two-way ANOVA: the treatment levels are fully orthogonal (all abundance levels of Species A are represented with all abundance levels of Species B). (See Inouye 2001 for more details. Figure modified from Goldberg and Scheiner 2001.)

authors have argued that such changes in density are indeed observed when a new species enters a community and establishes a population, so that adjusting for total density is not necessarily appropriate (Schluter 1995).

In a substitutive design, total density of organisms is kept constant, but the relative proportions of the two competitors are varied. These designs measure the relative intensity of inter- and intraspecific competition, but they do not measure the absolute strength of competition, and they assume responses are comparable at different density levels.

The response-surface design is a fully-crossed, two-way design that varies both the relative proportion and density of competitors. This design can be used to measure both relative intensity and absolute strength of inter- and intraspecific competitive interactions. However, as with all two-factor experiments with many treatment levels, adequate replication may be a problem. Inouye (2001) thoroughly reviews response-surface designs and other alternatives for competition studies.

Other issues that need to be addressed in competition experiments include: how many density levels to incorporate in order to estimate accurately competitive effects; how to deal with the non-independence of individuals within a treatment replicate; whether to manipulate or control for background species; and how to deal with residual carry-over effects and spatial heterogeneity that are generated by removal experiments, in which plots are established based on the presence of a species (Goldberg and Scheiner 2001).

SPLIT-PLOT DESIGNS The split-plot design is an extension of the randomized block design to two experimental treatments. The terminology comes from agricultural studies in which a single plot is split into subplots, each of which receives a different treatment. For our purposes, such a split-plot is equivalent to a block that contains within it different treatment replicates.

What distinguishes a split-plot design from a randomized block design is that a second treatment factor is also applied, this time at the level of the entire plot. Let's return one last time to the barnacle example. Once again, you are going to set up a two-way design, testing for predation and substrate effects. However, suppose that the cages are expensive and time-consuming to construct, and that you suspect there is a lot of microhabitat variation in the environment that is affecting your results. In a split-plot design, you would group the three substrates together, just as you did in the randomized block design. However, you would then place a *single cage* over all three of the substrate replicates within a single block. In this design, the predation treatment is referred to as the **whole-plot factor** because a single predation treatment is applied to an entire block. The substrate treatment is referred to as the **subplot factor** because all substrate treatments are applied within a single block. The split-plot design is illustrated in Figure 7.12.

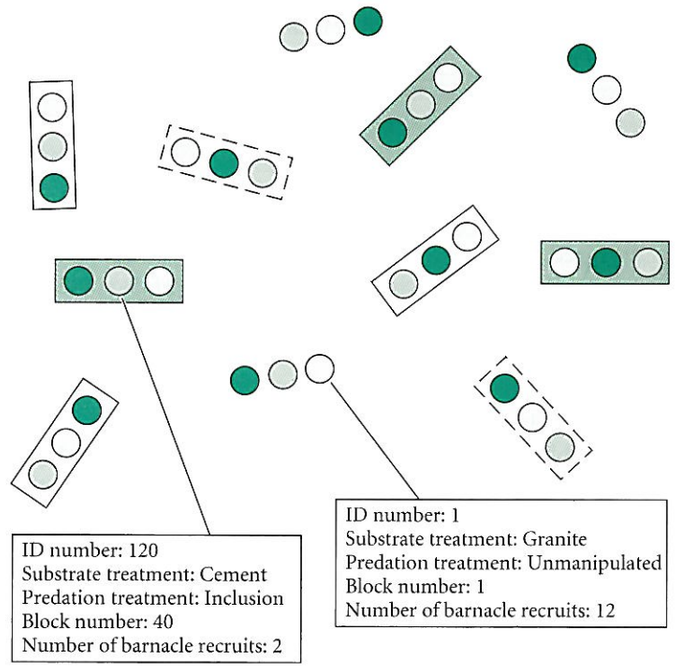


Figure 7.12 Example of a split-plot design. Treatment symbols are given in Figure 7.9. The three substrate treatments (subplot factor) are grouped in blocks. The predation treatment (whole plot factor) is applied to an entire block. The spreadsheet contains columns to indicate the substrate treatment, predation treatment, and block identity for each replicate. Only a subset of the blocks in each predation treatment are illustrated. The split-plot design is similar to a randomized block design (Figure 7.6), but in this case a second treatment factor is applied to the entire block (= plot).

ID number	Substrate treatment	Predation treatment	Block number	Number of barnacle recruits
1	Granite	Unmanipulated	1	12
2	Slate	Unmanipulated	1	10
3	Cement	Unmanipulated	1	8
4	Granite	Control	2	14
5	Slate	Control	2	10
6	Cement	Control	2	8
7	Granite	Predator exclusion	3	50
8	Slate	Predator exclusion	3	68
9	Cement	Predator exclusion	3	39
.	.	.	.	.
.	.	.	.	.
120	Cement	Predator inclusion	40	2

You should compare carefully the two-way layout (Figure 7.10), and the split-plot layout (Figure 7.12) and appreciate the subtle difference between them. The distinction is that, in the two-way layout, each replicate receives the treatment applications independently and separately. In the split-plot layout, one of the treatments is applied to entire blocks or plots, and the other treatment is applied to replicates within blocks.

The chief advantage of the split-plot design is the efficient use of blocks for the application of two treatments. As in the randomized block design, this is a simple layout that controls for environmental heterogeneity. It also may be less labor-intensive than applying treatments to individual replicates in a simple two-way design. The split-plot design removes the additive effects of the blocks and allows for tests of the main effects and interactions between the two manipulated factors.[7]

As in the randomized block design, the split-plot design does not allow you to test for the interaction between blocks and the subplot factor. However, the split-plot design does let you test for the main effect of the whole-plot factor, the main effect of the subplot factor, and the interaction between the two. As with nested designs, a very common mistake is for investigators to analyze a split-plot design as a two-factor ANOVA, which increases the risk of a Type I error.

DESIGNS FOR THREE OR MORE FACTORS The two-way design can be extended to three or even more factors. For example, if you were studying trophic cascades in a freshwater food web (Brett and Goldman 1997) you might add or remove top carnivores, predators, and herbivores, and then measure the effects on the producer level. This simple three-way design generates $2^3 = 8$ treatment combinations, including one combination that has neither top carnivores, predators, nor herbivores (Table 7.3). As we noted above, if you set up a randomized block design with a two-way layout and then replicate within blocks, the blocks then become a third factor in the analysis. However, three-factor (and higher) designs are used rarely in ecological studies. There are simply too many treatment combinations to make these designs logistically feasible. If you find

[7] Although the example we presented used two experimentally manipulated factors, the split-plot design is also effective when one of the two factors represents a source of natural variation. For example, in our research, we have studied the organization of aquatic food webs that develop in the rain-filled leaves of the pitcher plant *Sarracenia purpurea*. A new pitcher opens about once every 20 days, fills with rainwater, and quickly develops an associated food web of invertebrates and microorganisms.

In one of our experiments, we manipulated the disturbance regime by adding or removing water from the leaves of each plant (Gotelli et al., unpublished). These water manipulations were applied to all of the pitchers of a plant. Next, we recorded food web structure in the first, second, and third pitchers. These data are analyzed as a split-plot design. The whole-plot factor is water treatment (5 levels) and the subplot factor is pitcher age (3 levels). The plant served as a natural block, and it was efficient and realistic to apply the water treatments to all the leaves of a plant.

TABLE 7.3 **Treatment combinations in a three-way layout for a food web addition and removal experiment**

	Carnivore absent		Carnivore present	
	Herbivore absent	**Herbivore present**	**Herbivore absent**	**Herbivore present**
Producer absent	[10]	10	10	10
Producer present	10	10	10	10

In this experiment, the three trophic groups represent the three experimental factors (Carnivore, Herbivore, Producer), each of which has two levels (Absent, Present). The entry in each cell is the number of replicates of each treatment combination. If the response variable is some property of the food web itself, then the treatment combination in which all three trophic levels are absent (boxed) is not logically possible.

your design becoming too large and complex, you should consider breaking it down into a number of smaller experiments that address the key hypotheses you want to test.

INCORPORATING TEMPORAL VARIABILITY: REPEATED MEASURES DESIGNS In all of the designs we have described so far, the response variable is measured for each replicate at a single point in time at the end of the experiment. A **repeated measures design** is used whenever multiple observations on the same replicate are collected at different times. The repeated measures design can be thought of as a split-plot design in which a single replicate serves as a block, and the subplot factor is time. Repeated measures designs were first used in medical and psychological studies in which repeated observations were taken on an individual subject. Thus, in repeated measures terminology, the **between-subjects factor** corresponds to the whole-plot factor, and the **within-subjects factor** corresponds to the different times. In a repeated measures design, however, the multiple observations on a single individual are not independent of one another, and the analysis must proceed cautiously.

For example, suppose we used the simple one-way design for the barnacle study shown in Figure 7.5. But rather than censusing each replicate once, we measured the number of new barnacle recruits on each replicate for 4 consecutive weeks. Now, instead of 3 treatments × 10 replicates = 30 observations, we have 3 treatments × 10 replicates × 4 weeks = 120 observations (Table 7.4). If we only used data from one of the four censuses, the analysis would be identical to the one-way layout.

There are three advantages to a repeated measures design; the first is efficiency. Data are recorded at different times, but it is not necessary to have unique replicates for each time × treatment combination. Second, the repeated measures

TABLE 7.4 **Spreadsheet for a simple repeated measures analysis**

			Barnacle recruitment censuses			
ID number	**Treatment**	**Replicate**	**Week 1**	**Week 2**	**Week 3**	**Week 4**
1	Granite	1	12	15	17	17
2	Slate	1	10	6	19	32
3	Cement	1	3	2	0	2
4	Granite	2	14	14	5	11
5	Slate	2	10	11	13	15
6	Cement	2	8	9	4	4
7	Granite	3	11	13	22	29
8	Slate	3	11	17	28	15
9	Cement	3	7	7	7	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	Cement	10	8	0	0	3

This experiment is designed to test for the effect of substrate type on barnacle recruitment in the rocky intertidal (Figure 7.5). The data are organized in a spreadsheet in which each row is an independent replicate. The columns indicate the ID number (1–30), the treatment group (Cement, Slate, or Granite), and the replicate number (1–10 within each treatment). The next four columns give the number of barnacle recruits recorded on a particular substrate in each of four consecutive weeks. The measurements at different times are not independent of one another because they are taken from the same replicate each week.

design allows each replicate to serve as its own block or control. When the replicates represent individuals (plants, animals, or humans), this effectively controls for variation in size, age, and individual history, which often have strong influences on the response variable. Finally, the repeated measures design allows us to test for interactions of time with treatment. For many reasons, we expect that differences among treatments may change with time. In a press experiment (Chapter 6), there may be cumulative effects of the treatment that are not expressed until some time after the start of the experiment. In contrast, in a pulse experiment, we expect to see differences among treatments diminish as more time passes following the single, pulsed treatment application. Such complex effects are best seen in the interaction between time and treatment, and they may not be detected if the response variable is measured at only a single point in time.

Both the randomized block and the repeated measures designs make a special assumption of **circularity** for the within-subjects factor. Circularity (in the context of ANOVA) means that the variance of the difference between any two treatment levels in the subplot is the same. For the randomized block design, this means that the variance of the difference between any pair of treatments in the block is the same. If the treatment plots are large enough and adequately spaced, this is often a reasonable assumption. For the repeated measures design, the assumption of circularity means that the variance of the difference of observations between any pair of times is the same. This assumption of circularity is unlikely to be met for repeated measures; in most cases, the variance of the difference between two consecutive observations is likely to be much smaller than the variance of the difference between two observations that are widely separated in time. This is because time series measured on the same subject are likely to have a temporal "memory" such that current values are a function of values observed in the recent past. This premise of correlated observations is the basis for time-series analysis (Chapter 6).

The chief disadvantage with repeated measures analysis is failure to meet the assumption of circularity. If the repeated measures are serially correlated, Type I error rates for F-tests will be inflated, and the null hypothesis may be incorrectly rejected when it is true. The best way to meet the circularity assumption is to use evenly spaced sampling times along with knowledge of the natural history of your organisms to select an appropriate sampling interval.

What are some alternatives to repeated measures analysis that do not rely on the assumption of circularity? One approach is to set out enough replicates so that a different set is censused at each time period. With this design, time can be treated as a simple factor in a two-way analysis of variance. If the sampling methods are destructive (e.g., collecting stomach contents of fish, killing and preserving an invertebrate sample, or harvesting plants), this is the only method for incorporating time into the design.

A second strategy is to use the repeated measures layout, but to be more creative in the design of the response variable. Collapse the correlated repeated measures into a single response variable for each individual, and then use a simple one-way analysis of variance. For example, if you want to test whether temporal trends differ among the treatments (the between-subjects factor), you could fit a regression line (with either a linear or time-series model) to the repeated measures data, and use the slope of the line as the response variable. A separate slope value would be calculated for each of the individuals in the study. The slopes would then be compared using a simple one-way analysis, treating each individual as an independent observation (which it is). Significant treatment effects would indicate different temporal trajectories for individuals in the different treatments; this test is very similar to the test for an interaction of time and treatment in a standard repeated measures analysis.

Although such composite variables are created from correlated observations collected from one individual, they are independent among individuals. Moreover, the Central Limit Theorem (Chapter 2) tells us that averages of these

values will follow an approximately normal distribution even if the underlying variables themselves do not. Because most repeated measures data do not meet the assumption of circularity, our advice is to be careful with these analyses. We prefer to collapse the temporal data to a single variable that is truly independent among observations and then use a simpler one-way design for the analysis.

ENVIRONMENTAL IMPACTS OVER TIME: BACI DESIGNS A special type of repeated measures design is one in which measurements are taken both before and after the application of a treatment. For example, suppose you are interested in measuring the effect of atrazine (a hormone-mimicking compound) on the body size of frogs (Allran and Karasov 2001). In a simple one-way layout, you could assign frogs randomly to control and treatment groups, apply the atrazine, and measure body size at the end of the experiment. A more sensitive design might be to establish the control and treatment groups, and take measurements of body size for one or more time periods before application of the treatment. After the treatment is applied, you again measure body size at several times in both the control and treatment groups.

These designs also are used for observational studies assessing environmental impacts. In impact assessment, the measurements are taken before and after the impact occurs. A typical assessment might be the potential responses of a marine invertebrate community to the operation of a nuclear power plant, which discharges considerable hot water effluent (Schroeter et al. 1993). Before the power plant begins operation, you take one or more samples in the area that will be affected by the plant and estimate the abundance of species of interest (e.g., snails, sea stars, sea urchins). Replication in this study could be spatial, temporal, or both. Spatial replication would require sampling several different plots, both within and beyond the projected plume of hot water discharge.[8] Temporal replication would require sampling a single site in the discharge area at several times before the plant came on-line. Ideally, multiple sites would be sampled several times in the pre-discharge period.

Once the discharge begins, the sampling protocol is repeated. In this assessment design, it is imperative that there be at least one control or reference site that is sampled at the same time, both before and after the discharge. Then, if you observe a

[8] A key assumption in this layout is that the investigator knows ahead of time the spatial extent of the impact. Without this information, some of the "control" plots may end up within the "treatment" region, and the effect of the hot water plume would be underestimated.

decline in abundance at the impacted site but not at the control site, you can test whether the decline is significant. Alternatively, invertebrate abundance might be declining for reasons that have nothing to do with hot water discharge. In this case, you would find lower abundance at both the control and the impacted sites.

This kind of repeated measures design is referred to as a **BACI design** (Before-After, Control-Impact). Not only is there replication of control and treatment plots, there is temporal replication with measurements before and after treatment application (Figure 7.13).

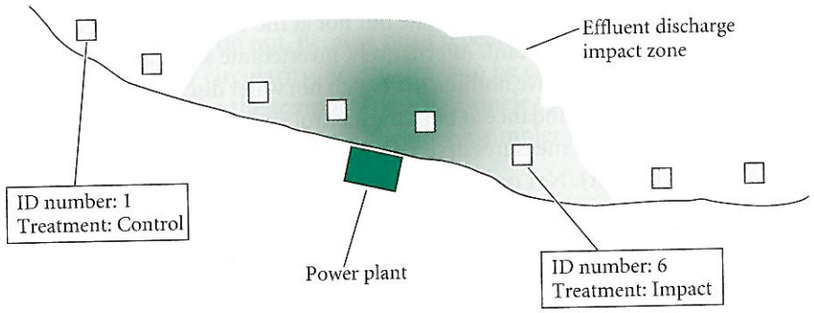
In its ideal form, the BACI design is a powerful layout for assessing environmental perturbations and monitoring trajectories before and after the impact. Replication in space ensures that the results will be applicable to other sites that may be perturbed in the same way. Replication through time ensures that the temporal trajectory of response and recovery can be monitored. The design is appropriate for both pulse and press experiments or disturbances.

Unfortunately, this idealized BACI design is rarely achieved in environmental impact studies. Often times, there is only a single site that will be impacted, and that site usually is not chosen randomly (Stewart-Oaten and Bence 2001, Murtaugh 2002b). Spatial replicates within the impacted area are not independent replicates because there is only a single impact studied at a single site (Underwood 1994). If the impact represents an environmental accident, such as an oil spill, no pre-impact data may be available, either from reference or impact sites.

The potential control of randomization and treatment assignments is much better in large-scale experimental manipulations, but even in these cases there may be little, if any, spatial replication. These studies rely on more intensive temporal replication, both before and after the manipulation. For example, since 1983, Brezonik et al. (1986) have conducted a long-term acidification experiment on Little Rock Lake, a small oligotrophic seepage lake in northern Wisconsin. The lake was divided into a treatment and reference basin with an impermeable vinyl curtain. Baseline (pre-manipulation) data were collected in both basins from August 1983 through April 1985. The treatment basin was then acidified with sulfuric acid in a stepwise fashion to three target pH levels (5.6, 5.1, 4.7). These pH levels were maintained in a press experiment at two-year intervals.

Because there is only one treatment basin and one control basin, conventional ANOVA methods cannot be used to analyze these data.[9] There are two general

[9] Of course, if one assumes the samples are independent replicates, conventional ANOVA could be used. But there is no wisdom in forcing data into a model structure they do not fit. One of the key themes of this chapter is to choose simple designs for your experiments and surveys whose assumptions best meet the constraints of your data.



ID number	Treatment	Pre-impact sampling				Post-impact sampling			
		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
1	Control	106	108	108	120	122	123	130	190
2	Control	104	88	84	104	106	119	135	120
3	Impact	99	97	102	192	150	140	145	150
4	Impact	120	122	98	120	137	135	155	165
5	Impact	88	90	92	94	0	7	75	77
6	Impact	100	120	129	82	2	3	66	130
7	Control	66	70	70	99	45	55	55	109
8	Control	130	209	220	250	100	90	88	140

Figure 7.13 Example of a spatial arrangement of replicates for a BACI design. Each square represents a sample plot on a shoreline that will potentially be affected by hot water discharged from a nuclear power plant. Permanent plots are established within the hot water effluent zone (shaded area), and in adjacent control zones (unshaded areas). All plots are sampled weekly for 4 weeks before the plant begins discharging hot water and for 4 weeks afterwards. Each row of the spreadsheet represents a different replicate. Two columns indicate the replicate ID number and the treatment (Control or Impact). The remaining columns give the invertebrate abundance data collected at each of the 8 sampling dates (4 sample dates pre-discharge, 4 sample dates post-discharge).

strategies for analysis. **Randomized intervention analysis** (RIA) is a Monte Carlo procedure (see Chapter 5) in which a test statistic calculated from the time series is compared to a distribution of values created by randomizing or reshuffling the time-series data among the treatment intervals (Carpenter et al. 1989). RIA relaxes the assumption of normality, but it is still susceptible to temporal correlations in the data (Stewart-Oaten et al. 1992).

A second strategy is to use time series analysis to fit simple models to the data. The **autoregressive integrated moving average** (ARIMA) model describes the correlation structure in temporal data with a few parameters (see Chapter 6). Additional model parameters estimate the stepwise changes that occur with the experimental interventions, and these parameters can then be tested against the

null hypothesis that they do not differ from 0. ARIMA models can be fit individually to the control and manipulation time series data, or to a derived data series created by taking the ratio of the treatment/control data at each time step (Rasmussen et al. 1993). Bayesian methods can also be used to analyze data from BACI designs (Carpenter et al. 1996; Rao and Tirtotjondro 1996; Reckhow 1996; Varis and Kuikka 1997; Fox 2001).

RIA, ARIMA, and Bayesian methods are powerful tools for detecting treatment effects in time-series data. However, without replication, it is still problematic to generalize the results of the analysis. What might happen in other lakes? Other years? Additional information, including the results of small-scale experiments (Frost et al. 1988), or snapshot comparisons with a large number of unmanipulated control sites (Schindler et al. 1985, Underwood 1994) can help to expand the domain of inference from BACI study.

Alternatives to ANOVA: Experimental Regression

The literature on experimental design is dominated by ANOVA layouts, and modern ecological science has been referred to sarcastically as little more than the care and curation of ANOVA tables. Although ANOVA designs are convenient and powerful for many purposes, they are not always the best choice. ANOVA has become so popular that it may act as an intellectual straightjacket (Werner 1998), and cause scientists to neglect other useful experimental designs.

We suggest that ANOVA designs often are employed when a regression design would be more appropriate. In many ANOVA designs, a continuous predictor variable is tested at only a few values so that it can be treated as a categorical predictor variable, and shoehorned into an ANOVA design. Examples include treatment levels that represent different nutrient concentrations, temperatures, or resource levels.

In contrast, an experimental regression design (Figure 7.14) uses many different levels of the continuous independent variable, and then uses regression to fit a line, curve, or surface to the data. One tricky issue in such a design (the same problem is present in an ANOVA) is to choose appropriate levels for the predictor variable. A uniform selection of predictor values within the desired range should ensure high statistical power and a reasonable fit to the regression line. However, if the response is expected to be multiplicative rather than linear (e.g., a 10% decrease in growth with every doubling of concentration), it might be better to set the predictor values on an evenly spaced logarithmic scale. In this design, you will have more data collected at low concentrations, where changes in the response variable might be expected to be steepest.

One of the chief advantages of regression designs is efficiency. Suppose you are studying responses of terrestrial plant and insect communities to nitrogen

(Simlutaneous N and P treatments in a two-way ANOVA design)

		Nitrogen treatment	
		0.00 mg	0.50 mg
Phosphorus treatment	0.00 mg	12	12
	0.05 mg	12	12

(Two-way experimental regression design)

		Nitrogen treatment						
		0.00 mg	0.05 mg	0.10 mg	0.20 mg	0.40 mg	0.80 mg	1.00 mg
Phosphorus treatment	0.00 mg	1	1	1	1	1	1	1
	0.01 mg	1	1	1	1	1	1	1
	0.05 mg	1	1	1	1	1	1	1
	0.10 mg	1	1	1	1	1	1	1
	0.20 mg	1	1	1	1	1	1	1
	0.40 mg	1	1	1	1	1	1	1
	0.50 mg	1	1	1	1	1	1	1

Figure 7.14 Treatment combinations in a two-way ANOVA design (upper panel) and an experimental regression design (lower panel). These experiments test for additive and interactive effects of nitrogen (N) and phosphorus (P) on plant growth or some other response variable. Each cell in the table indicates the number of replicate plots used. If the maximum number of replicates is 50 and a minimum of 10 replicates per treatment in required, only 2 treatment levels each for N and P are possible (each with 12 replicates) in the two-way ANOVA. In contrast, the experimental regression allows for 7 treatment levels each of N and P. Each of the $7 \times 7 = 49$ plots in the design receives a unique combination of N and P concentrations.

(N), and your total sample size is limited by available space or labor to 50 plots. If you try to follow the Rule of 10, an ANOVA design would force you to select only 5 different fertilization levels and replicate each one 10 times. Although this design is adequate for some purposes, it may not help to pinpoint critical threshold levels at which community structure changes dramatically in response to N. In contrast, a regression design would allow you to set up 50 different N levels, one in each of the plots. With this design, you could very accurately characterize changes that occur in community structure with increasing N levels; graphical displays may help to reveal threshold points and non-linear effects (see Chapter 9). Of course, even minimal replication of each treatment level is very desirable, but if the total sample size is limited, this may not be possible.

For a two-factor ANOVA, the experimental regression is even more efficient and powerful. If you want to manipulate nitrogen and phosphorus (P) as independent factors, and still maintain 10 replicates per treatment, you could have no more than 2 levels of N and 2 levels of P. Because one of those levels must be a control plot (i.e., no fertilization), the experiment isn't going to give you very much information about the role of changing levels of N and P on the

system. If the result is statistically significant, you can say only that the community responds to those particular levels of N and P, which is something that you may have already known from the literature before you started. If the result is not statistically significant, the obvious criticism would be that the concentrations of N and P were too low to generate a response.

In contrast, an experimental regression design would be a fully crossed design with 7 levels of nitrogen and 7 levels of phosphorus, with one level of each corresponding to the control (no N or P). The $7 \times 7 = 49$ replicates each receive a unique concentration of N and P, with one of the 49 plots receiving neither N nor P (Figure 7.14). This is a response surface design (Inouye 2001), in which the response variable will be modeled by multiple regression. With seven levels of each nutrient, this design provides a much more powerful test for additive and interactive effects of nutrients, and could also reveal non-linear responses. If the effects of N and P are weak or subtle, the regression model will be more likely to reveal significant effects than the two-way ANOVA.[10]

Efficiency is not the only advantage of an experimental regression design. By representing the predictor variable naturally on a continuous scale, it is much easier to detect non-linear, threshold, or asymptotic responses. These cannot be inferred reliably from an ANOVA design, which usually will not have enough treatment levels to be informative. If the relationship is something other than a straight line, there are a number of statistical methods for fitting non-linear responses (see Chapter 9).

A final advantage to using an experimental regression design is the potential benefits for integrating your results with theoretical predictions and ecological models. ANOVA provides estimates of means and variances for groups or particular levels of categorical variables. These estimates are rarely of interest or use in ecological models. In contrast, a regression analysis provides esti-

[10] There is a further hidden penalty in using the two-way ANOVA design for this experiment that often is not appreciated. If the treatment levels represent a small subset of many possible other levels that could have been used, then the design is referred to as a **random effects ANOVA** model. Unless there is something special about the particular treatment levels that were used, a random effects model is always the most appropriate choice when a continuous variable has been shoehorned into a categorical variable for ANOVA. In the random effects model, the denominator of the F-ratio test for treatment effects is the interaction sum of squares, not the error sum of squares that is used in a standard **fixed effects ANOVA** model. If there are not many treatment levels, there will not be very many degrees of freedom associated with the interaction term, regardless of the amount of replication within treatments. As a consequence, the test will be much less powerful than a typical fixed effects ANOVA. See Chapter 10 for more details on fixed and random effects ANOVA models and the construction of F-ratios.

mates of slope and intercept parameters that measure the change in the response Y relative to a change in predictor X (dY/dX). These derivatives are precisely what are needed for testing the many ecological models that are written as simple differential equations.

An experimental regression approach might not be feasible if it is very expensive or time-consuming to establish unique levels of the predictor variable. In that case, an ANOVA design may be preferred because only a few levels of the predictor variable can be established. An apparent disadvantage of the experimental regression is that it appears to have no replication! In Figure 7.14, each unique treatment level is applied to only a single replicate, and that seems to fly in the face of the principle of replication (see Chapter 6). If each unique treatment is unreplicated, the least-squares solution to the regression line still provides an estimate of the regression parameters and their variances (see Chapter 9). The regression line provides an unbiased estimate of the expected value of the response Y for a given value of the predictor X, and the variance estimates can be used to construct a confidence interval about that expectation. This is actually more informative than the results of an ANOVA model, which allow you to estimate means and confidence intervals for only the handful of treatment levels that were used.

A potentially more serious issue is that the regression design may not include any replication of controls. In our two-factor example, there is only one plot that contains no nitrogen and no phosphorus addition. Whether this is a serious problem or not depends on the details of the experimental design. As long as all of the replicates are treated the same and differ only in the treatment application, the experiment is still a valid one, and the results will estimate accurately the relative effects of different levels of the predictor variable on the response variable. If it is desirable to estimate the absolute treatment effect, then additional replicated control plots may be needed to account for any handling effects or other responses to the general experimental conditions. These issues are no different than those that are encountered in ANOVA designs.

Historically, regression has been used predominantly in the analysis of nonexperimental data, even though its assumptions are unlikely to be met in most sampling studies. An experimental study based on a regression design not only meets the assumptions of the analysis, but is often more powerful and appropriate than an ANOVA design. We encourage you to "think outside the ANOVA box" and consider a regression design when you are manipulating a continuous predictor variable.

Tabular Designs

The last class of experimental designs is used when both predictor and response variables are categorical. The measurements in these designs are counts. The simplest such variable is a dichotomous (or binomial, see Chapter 2) response in a series of independent trials. For example, in a test of cockroach behavior, you could place an individual cockroach in an arena with a black and a white side, and then record on which side the animal spent the majority of its time. To ensure independence, each replicate cockroach would be tested individually.

More typically, a dichotomous response will be recorded for two or more categories of the predictor variable. In the cockroach study, half of the cockroaches might be infected experimentally with a parasite that is known to alter host behavior (Moore 1984). Now we want to ask whether the response of the cockroach differs between parasitized and unparasitized individuals (Moore 2001; Poulin 2000). This approach could be extended to a three-way design by adding an additional treatment and asking whether the difference between parasitized and unparasitized individuals changes in the presence or absence of a vertebrate predator.

We might predict that uninfected individuals are more likely to use the black substrate, which will make them less conspicuous to a visual predator. In the presence of a predator, uninfected individuals might shift even more towards the dark surfaces, whereas infected individuals might shift more towards white surfaces. Alternatively, the parasite might alter host behavior, but those alterations might be independent of the presence or absence of the predator. Still another possibility is that host behavior might be very sensitive to the presence of the predator, but not necessarily affected by parasite infection. A **contingency table analysis** (Chapter 11) is used to test all these hypotheses with the same dataset.

In some tabular designs, the investigator determines the total number of individuals in each category of predictor variable, and these individuals will be classified according to their responses. The total for each category is referred to as the **marginal total** because it represents the column or row sum in the margin of the data table. In an observational study, the investigator might determine one or both of the marginal totals, or perhaps only the grand total of independent observations. In a tabular design, the grand total equals the sum of either the column or row marginal totals.

For example, suppose you are trying to determine the associations of four species of *Anolis* lizard with three microhabitat types (Ground, Tree trunks, Tree branches; see Butler and Losos 2002). Table 7.5 shows the two-way layout of the data from such a study. Each row in the table represents a different lizard species, and each column represents a different habitat category. The entries in each cell represent the counts of a particular lizard species recorded in a particular habitat. The marginal row totals represent the total number of observations for each lizard species, summed across the three habitat types. The marginal column totals

TABLE 7.5 **Tabulated counts of the occurrence of four lizard species censused in three different microhabitats**

		Habitat			
		Ground	Tree trunk	Tree branch	Species totals
Lizard species	Species A	9	0	15	*24*
	Species B	9	0	12	*21*
	Species C	9	5	0	*14*
	Species D	9	10	3	*22*
	Habitat Totals	*36*	*15*	*30*	*81*

Italicized values are the marginal totals for the two-way table. The total sample size is 81 observations. In these data, both the response variable (species identity) and the predictor variable (microhabitat category) are categorical.

represent the total number of observations in each habitat type, summed across the three lizard species. The grand total in the table ($N = 81$) represents the total count of all lizards species observed in all habitats.

There are several ways that these data could have been collected, depending on whether the sampling was based on the marginal totals for the microhabitats, the marginal totals for the lizards, or the grand total for the entire sample.

In a sampling scheme built around the microhabitats, the investigator might have spent 10 hours sampling each microhabitat, and recording the number of different lizard species encountered in each habitat. Thus, in the census of tree trunks, the investigator found a total of 15 lizards: 5 of Species C and 10 of Species D; Species A and B were never encountered. In the ground census, the investigator found a total of 36 lizards, with all four species equally represented.

Alternatively, the sampling could have been based on the lizards themselves. In such a design, the investigator would put in an equal sampling effort for each species by searching all habitats randomly for individuals of a particular species of lizard and then recording in which microhabitat they occurred. Thus, in a search for Species B, 21 individuals were found, 9 on the ground, and 12 on tree branches. Another sampling variant is one in which the row and column totals are simultaneously fixed. Although this design is not very common in ecological studies, it can be used for an exact statistical test of the distribution of sampled values (Fisher's Exact Test; see Chapter 11). Finally, the sampling might have been based simply on the grand total of observations. Thus, the investigator might have taken a random sample of 81 lizards, and for each lizard encountered, recorded the species identity and the microhabitat.

Ideally, the marginal totals on which the sampling is based should be the same for each category, just as we try to achieve equal sample sizes in setting up an

ANOVA design. However, identical sample sizes are not necessary for analysis of tabular data. However, the tests do require (as always) that the observations be randomly sampled and that the replicates be truly independent of one another. This may be very difficult to achieve in some cases. For example, if lizards tend to aggregate or move in groups, we cannot simply count individuals as they are encountered, because an entire group is likely to be found in a single microhabitat. In this example, we also assume that all of the lizards are equally conspicuous to the observer in all of the habitats. If some species are more obvious in certain habitats than others, then the relative frequencies will reflect sampling biases rather than species' microhabitat associations.

SAMPLING DESIGNS FOR CATEGORICAL DATA In contrast to the large literature for regression and ANOVA designs, relatively little has been written, in an ecological context, about sampling designs for categorical data. If the observations are expensive or time-consuming, every effort should be made to ensure that each observation is independent, so that a simple two- or multiway layout can be used. Unfortunately, many published analyses of categorical data are based on nonindependent observations, some of it collected in different times or in different places. Many behavioral studies analyze multiple observations of the same individual. Such data clearly should not be treated as independent (Kramer and Schmidhammer 1992). If the tabular data are not independent, random samples from the same sampling space, you should explicitly incorporate the temporal or spatial categories as factors in your analysis.

Alternatives to Tabular Designs: Proportional Designs

If the individual observations are inexpensive and can be gathered in large numbers, there is an alternative to tabular designs. One of the categorical variables can be collapsed to a measure of proportion (number of desired outcomes/number of observations), which is a continuous variable. The continuous variable can then be analyzed using any of the methods described above for regression or ANOVA.

There are two advantages of using proportional designs in lieu of tabular ones. The first is that the standard set of ANOVA and regression designs can be used, including blocking. The second advantage is that the analysis of proportions can be used to accommodate frequency data that are not strictly independent. For example, suppose that, to save time, the cockroach experiment were set up with 10 individuals placed in the behavioral arena at one time. It would not be legitimate to treat the 10 individuals as independent replicates, for the same reason that subsamples from a single cage are not independent replicates for an ANOVA. However, the data from this run can be treated as a single repli-

cate, for which we could calculate the proportion of individuals present on the black side of the arena. With multiple runs of the experiment, we can now test hypotheses about differences in the proportion among groups (e.g., parasitized versus unparasitized). The design is still problematic, because it is possible that substrate selection by solitary cockroaches may be different from substrate selection by groups of cockroaches. Nevertheless, the design at least avoids treating individuals within an arena as independent replicates, which they are not.

Although proportions, like probabilities, are continuous variables, they are bounded between 0.0 and 1.0. An arcsin square root transformation of proportional data may be necessary to meet the assumption of normality (Chapter 8). A second consideration in the analysis of proportions is that it is very important to use at least 10 trials per replicate, and to make sure that sample sizes are as closely balanced as possible. With 10 individuals per replicate, the possible measures for the response variable are in the set {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. But suppose the same treatment is applied to a replicate in which only three individuals were used. In this case, the only possible values are in the set {0.0, 0.33, 0.66, 1.0}. These small sample sizes will greatly inflate the measured variance, and this problem is not alleviated by any data transformation.

A final problem with the analysis of proportions arises if there are three or more categorical variables. With a dichotomous response, the proportion completely characterizes the data. However, if there are more than two categories, the proportion will have to be carefully defined in terms of only one of the categories. For example, if the arena includes vertical and horizontal black and white surfaces, there are now four categories from which the proportion can be measured. Thus, the analysis might be based on the proportion of individuals using the horizontal black surface. Alternatively, the proportion could be defined in terms of two or more summed categories, such as the proportion of individuals using any vertical surface (white or black).

Summary

Independent and dependent variables are either categorical or continuous, and most designs fit into one of four possible categories based on this classification. Analysis of variance (ANOVA) designs are used for experiments in which the independent variable is categorical and the dependent variable is continuous. Useful ANOVA designs include one- and two-way ANOVAs, randomized block, and split-plot designs. We do not favor the use of nested ANOVAs, in which non-independent subsamples are taken from within a replicate. Repeated measures designs can be used when repeated observations are collected on a single replicate through time. However, these data are often autocorrelated, so

that the assumptions of the analysis may not be met. In such cases, the temporal data should be collapsed to a single independent measurement, or time-series analysis should be employed.

If the independent variable is continuous, a regression design is used. Regression designs are appropriate for both experimental and sampling studies, although they are used predominantly in the latter. We advocate increased use of experimental regression in lieu of ANOVA with only a few levels of the independent variable represented. Adequate sampling of the range of predictor values is important in designing a sound regression experiment. Multiple regression designs include two or more predictor variables, although the analysis becomes problematic if there are strong correlations (collinearity) among the predictor variables.

If both the independent and the dependent variables are categorical, a tabular design is employed. Tabular designs require true independence of the replicate counts. If the counts are not independent, they should be collapsed so that the response variable is a single proportion. The experimental design is then similar to a regression or ANOVA.

We favor simple experimental and sampling designs and emphasize the importance of collecting data from replicates that are independent of one another. Good replication and balanced sample sizes will improve the power and reliability of the analyses. Even the most sophisticated analysis cannot salvage results from a poorly designed study.