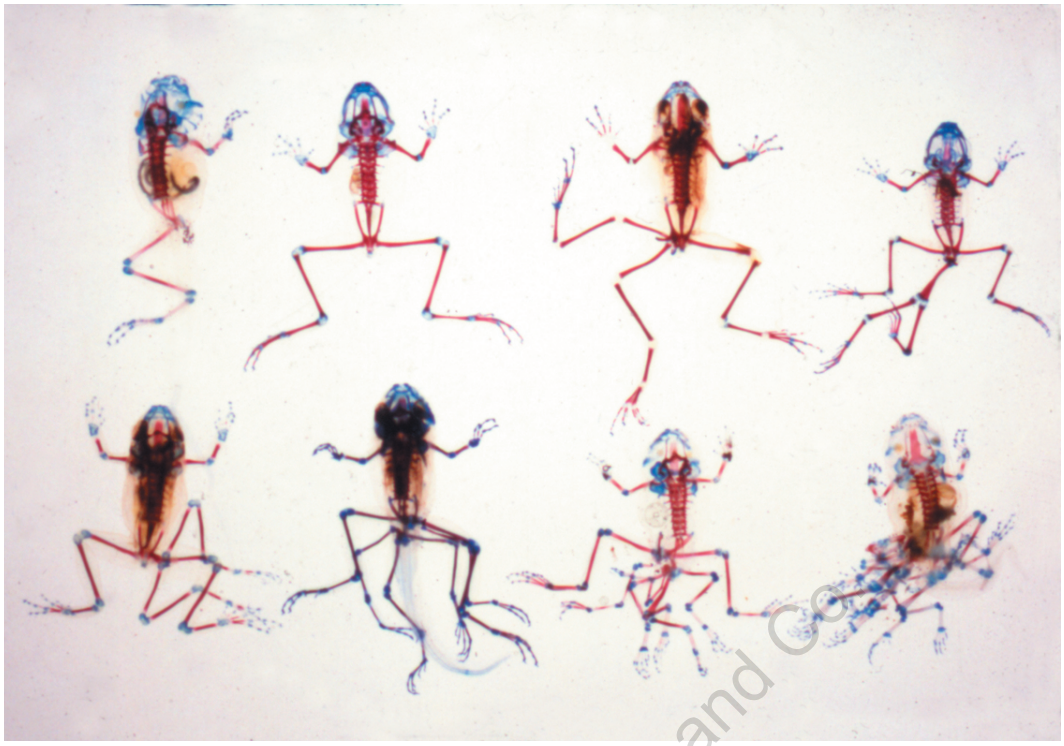

Frog deformities

14 Designing experiments

Two types of investigations are carried out in biology: observational and experimental. In an experimental study, the researcher assigns treatments to units or subjects so that differences in response can be compared. In an observational study, on the other hand, nature does the assigning of treatments to subjects. The researcher has no influence over which subjects receive which treatment.

What's so important about the distinction? Whereas observational studies can identify associations between treatment and response variables, properly designed experimental studies can identify the *causes* of these associations.

How do we best design an experiment to get the most information possible out of it? The short answer is that we must design to eliminate bias and to reduce the influence of sampling error. The present chapter outlines the basics on how to accomplish this feat. We also briefly discuss how to design an observational study: by taking the best features of experimental designs and incorporating as many of them as possible.

Finally, we discuss how to plan the sample size needed in an experimental or observational study.

14.1 Why do experiments?

In an experimental study, there must be at least two treatments and the experimenter (rather than nature) must assign them to units or subjects. The crucial advantage of experiments derives from the *random* assignment of treatments to units. Random assignment, or randomization, minimizes the influence of *confounding* variables (Interleaf 4), allowing the experimenter to isolate the effects of the treatment variable.

Confounding variables

Studies in biology are usually carried out with the aim of deciding how an explanatory variable or treatment affects a response variable. How are injury rates in cats with "high-rise syndrome" affected by the number of stories fallen? What is the effect of marine reserves on fish biomass? How does the use of supplemental oxygen affect the probability of surviving an ascent of Mount Everest? The easiest way to address these questions is with an observational study—that is, to gather measurements of both variables of interest on a set of subjects and estimate the association between them. If the two variables are correlated or associated, then one may be the cause of the other.

The limitation of the observational approach is that, by itself, it cannot distinguish between two completely different reasons behind an association between an explanatory variable X and a response variable Y. One possibility is that X really does cause a response in Y. For example, taking supplemental oxygen might increase the chance of survival during a climb of Mount Everest. The other possibility is that the explanatory variable X has no effect at all on the response variable Y; they are associated only because other variables affect both X and Y at the same time. For example, the use of supplemental oxygen might just be a benign indicator of a greater overall preparedness of the climbers who use it, and greater preparedness rather than oxygen use is the real cause of the enhanced survival. Variables (like preparedness) that distort the causal relationship between the measured variables of interest (oxygen use and survival) are called *confounding variables*. Recall from Interleaf 4, for example, that ice cream consumption and violent crime are correlated, but neither is the cause of the other. Instead, increases in both ice cream consumption and crime are caused by higher temperatures. Temperature is a confounding variable in this example.

A *confounding variable* is a variable that masks or distorts the causal relationship between measured variables in a study.

Confounding variables bias the estimate of the causal relationship between measured explanatory and response variables, sometimes even reversing the apparent effect of one on the other. For example, observational studies have indicated that breast-fed babies have lower weight at six and 12 months of age compared with formula-fed infants (Interleaf 4). But an experimental study using randomization found that mean infant weight was actually *higher* in breast-fed babies at six months of age and was not less than that in formula-fed babies at 12 months (Kramer et al. 2002). The observed relationship between breast feeding and infant growth was confounded by unmeasured variables such as the socioeconomic status of the parents.

With an experiment, random assignment of treatments to participants allows researchers to tease apart the effects of the explanatory variable from those of confounding variables. With random assignment, no confounding variables will be associated with treatment except by chance. For example, if women who choose to breast-feed their babies have a different average socioeconomic background than women who choose to feed their infants formula, randomly assigning the treatments "breast feeding" and "formula feeding" to women in an experiment will break this connection, roughly equalizing the backgrounds of the two treatment groups. In this case, any resulting difference between groups in infant weight (beyond chance) must be caused by treatment.

Experimental artifacts

Unfortunately, experiments themselves might inadvertently create artificial conditions that distort cause and effect. Experiments should be designed to minimize artifacts.

An *experimental artifact* is a bias in a measurement produced by unintended consequences of experimental procedures.

For example, experiments conducted on aquatic birds have shown that their heart rates drop sharply when they are forcibly submerged in water, compared with individuals remaining above water. The drop in heart rate has been interpreted as an oxygen-saving response. Later studies using improved technology showed that voluntary dives do not produce such a large drop in heart rate (Kanwisher et al. 1981). This finding suggested that a component of the heart rate response in forced dives was induced by the stress of being forcibly dunked underwater, rather than the dive itself. The experimental conditions introduced an artifact that for a while went unrecognized.

To prevent artifacts, experimental studies should be conducted under conditions that are as natural as possible. A potential drawback is that more natural conditions might introduce more sources of variation, reducing power and precision. Observational studies can provide important insight into what is the best setting for an experiment.

14.2 Lessons from clinical trials

The gold standard of experimental designs is the **clinical trial**, an experimental study in which two or more treatments are assigned to human participants. The design of clinical trials has been refined because the cost of making a mistake with human participants is so high. Experiments on nonhuman subjects are simply called "laboratory experiments" or "field experiments," depending on where they take place. Experimental studies in all areas of biology have been greatly informed by procedures used in clinical trials.

A *clinical trial* is an experimental study in which two or more treatments are applied to human participants.

Before we dig into the logic of the main components of experimental design, let's look at the clinical trial in Example 14.2, which incorporates many of these features.

EXAMPLE 14.2 Reducing HIV transmission

Transmission of the HIV-1 virus via sex workers contributes to the rapid spread of AIDS in Africa. How can this transmission be reduced? In laboratory experiments, the spermicide nonoxynol-9 had shown in vitro activity against HIV-1, shown schematically at the right. This finding motivated a clinical trial by van Damme et al. (2002), who tested whether a vaginal gel containing the chemical would reduce female sex workers' risk of acquiring the disease. Data were gathered on a volunteer sample of 765 HIV-free sex workers in six clinics in Asia and Africa. Two gel treatments were assigned randomly to women at each clinic. One gel contained nonoxynol-9, and the other contained a placebo (an inactive compound that participants could not distinguish from the treatment of interest). Neither the participants nor the researchers making observations at the clinics knew who had received the treatment and who had received the placebo. (A system of numbered codes kept track of who got which treatment.) By the end of the experiment, 59 of 376 women in the nonoxynol-9 group (15.9 %) were HIV-positive (Table 14.2-1), compared with 45 out of 389 women in the placebo group (11.6 %). Thus, the odds of contracting HIV-1 were slightly higher in the nonoxynol-9 group compared with the placebo group—which was the opposite of the expected result. The reason seems to be that repeated use of nonoxynol-9 causes tissue damage that leads to higher risk.

Design components

A good experiment is designed with two objectives in mind:

- To reduce bias in estimating and testing treatment effects
- To reduce the effects of sampling error

TABLE 14.2-1 Results of the clinical trial in Example 14.2 (*n* is the number of subjects).

	Nonoxynol-9		Placebo	
Clinic	*n*	Number infected	*n*	Number infected
Abidjan	78	0	84	5
Bangkok	26	0	25	0
Cotonou	100	12	103	10
Durban	94	42	93	30
Hat Yai 2	22	0	25	0
Hat Yai 3	56	5	59	0
Total	376	59	389	45

The most significant elements in the design of the clinical trial in Example 14.2 addressed these two objectives. To reduce bias, the experiment included the following elements.

1. A simultaneous control group: the study included both the treatment of interest and a control group (the women receiving the placebo).
2. Randomization: treatments were randomly assigned to women at each clinic.
3. Blinding: neither the participants nor the clinicians knew which women were assigned which treatment.

To reduce the effects of sampling error, the experiment included these elements.

1. Replication: the study was carried out on multiple independent participants.
2. Balance: the number of women was nearly equal in the two groups at every clinic.
3. Blocking: participants were grouped according to the clinic they attended, yielding multiple repetitions of the same experiment in different settings (i.e., "blocks").

The goal of experimental design is to eliminate bias and to reduce sampling error when estimating and testing the effects of one variable on another.

In Section 14.3, we discuss the virtues of the three main strategies used to reduce bias—namely, simultaneous controls, randomization, and blinding. In Section 14.4, we explain the strategies used to reduce the effects of sampling error—namely, replication, balance, and blocking. As usual, we assume throughout that units or subjects have been randomly sampled from the population of interest.

14.3 How to reduce bias

We have seen how confounding variables in observational studies can bias the estimated effects of an explanatory variable on a response variable. The following experimental procedures are meant to eliminate bias.

Simultaneous control group

A **control group** is a group of subjects who are treated like all of the experimental subjects, except that the control group does not receive the treatment of interest.

A *control group* is a group of subjects who do not receive the treatment of interest but who otherwise experience similar conditions as the treated subjects.

In an uncontrolled experiment, a group of subjects are treated in some way and then measured to see how they have responded. Lacking a control group for comparison, such a study cannot determine whether the treatment of interest is the cause of any of the observed changes. There are several possible reasons for this, including the following:

- Sick human participants selected for a medical treatment may tend to "bounce back" toward their average condition regardless of any effect of the treatment (Interleaf 6).
- Stress and other impacts associated with administering the treatment (such as surgery or confinement) might themselves produce a response separate from the effect of the treatment of interest.
- The health of human participants often improves after treatment merely because of their expectation that the treatment will have an effect. This phenomenon is known as the placebo effect (Interleaf 6).

The solution to all of these problems is to include a control group of subjects measured for comparison. The treatment and control subjects should be tested simultaneously or in random order, to ensure that any temporal changes in experimental conditions do not affect the outcome.

The appropriate control group will depend on the circumstance. Here are some examples:

- In clinical trials, either a placebo or the currently accepted treatment should be provided, such as in Example 14.2. A placebo is an inactive treatment that subjects cannot distinguish from the main treatment of interest.
- In experiments requiring intrusive methods to administer treatment, such as injections, surgery, restraint, or confinement, the control subjects should be

perturbed in the same way as the other subjects, except for the treatment itself, as far as ethical considerations permit. The "sham operation," in which surgery is carried out without the experimental treatment itself, is an example. Sham operations are very rare in human studies, but they are more common in animal experiments.

- In field experiments, applying a treatment of interest may physically disturb the plots receiving it and the surrounding areas, perhaps by the researchers trampling. Ideally, the same disturbance should be applied to the control plots.

Often it is desirable to have more than one control group. For example, two control groups, where one is a harmless placebo and the other is the best existing treatment, may be used in a study so that the total effect of the treatment and the improvement of the new treatment over the old may both be measured. However, using resources for multiple controls might reduce the power of the study to test its main hypotheses.

Randomization

Once the treatments have been chosen, the researcher should *randomize* their assignment to units or subjects in the sample. **Randomization** means that treatments are assigned to units at random, such as by flipping a coin. Chance rather than conscious or unconscious decision determines which units end up receiving the treatment of interest and which receive the control. A **completely randomized design** is an experimental design in which treatments are assigned to all units by randomization.

Randomization is the random assignment of treatments to units in an experimental study.

The virtue of randomization is that it breaks the association between possible confounding variables and the explanatory variable, allowing the causal relationship between the explanatory and response variables to be assessed. Randomization doesn't eliminate the variation contributed by confounding variables, only their correlation with treatment. It ensures that variation from confounding variables is spread more evenly between the different treatment groups, and so it creates no bias. If randomization is done properly, any remaining influence of confounding variables occurs by chance alone, which statistical methods can account for.

Randomization should be carried out using a random process. The following steps describe one way to assign treatments randomly:

1. List all n subjects, one per row, in a computer spreadsheet.
2. Use the computer to give each individual a random number.[1]
3. Assign treatment A to those subjects receiving the lowest numbers and treatment B to those with the highest numbers.

1. The Random.org website at http://random.org/sequences will also do this.

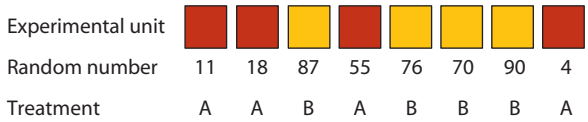

FIGURE 14.3-1 A procedure for randomization. Each of eight subjects was assigned a number between 0 and 99 that was drawn at random by a computer. Treatment A (*colored red*) was assigned to the four subjects with the lowest random numbers, whereas treatment B (*gold*) was assigned to the rest.

This process is demonstrated in Figure 14.3-1, where eight subjects are assigned to two treatments, A and B.

Other ways of assigning treatments to subjects are almost always inferior, because they do not eliminate the effects of confounding variables. For example, the following methods can lead to problems:

- Assign treatment A to all patients attending one clinic and treatment B to patients attending a second clinic. (Problem: All of the other differences between the two clinics become confounding variables. If one clinic is better than the other in general, then the difference in clinic quality would show up as a difference in treatments.)
- Assign treatments to human participants alphabetically. (Problem: This might inadvertently group individuals having the same nationality, generating unwanted differences between treatments in health histories and genetic variables.)

It is important to use a computer random-number generator or random-number tables to assign individuals randomly to treatments. "Haphazard" assignment, in which the researcher chooses a treatment while trying to make it random, has repeatedly been shown to be non-random and prone to bias.[2]

Blinding

The process of concealing information from participants and researchers about which of them receive which treatments is called **blinding**. Blinding prevents participants and researchers from changing their behavior, consciously or unconsciously, based on their knowledge of which treatment they were receiving or administering. For example, a researcher who believes that acupuncture helps alleviate back pain might unconsciously interpret a patient's report of pain differently if the researcher knows the patient was assigned the acupuncture treatment instead of a placebo. This might explain why studies that have shown acupuncture has a significant effect on back pain

2. What do you do if, by chance, the first four of eight units are all assigned treatment A and the last four are assigned treatment B, yielding the arrangement AAAABBBB? Some biologists might randomize again to ensure the interspersion of treatments, but that is not strictly legitimate. If the first four units are different somehow from the last four, apart from treatment, then blocking (Section 14.4) should be considered as a remedy.

are limited to those without blinding (Ernst and White 1998). Studies implementing blinding have not found that acupuncture has an ameliorating effect on back pain.

In a **single-blind experiment**, participants are unaware of the treatment they have been assigned. This requires that the treatments be indistinguishable to the participants, a particular necessity in experiments involving humans. Single-blinding prevents participants from responding differently according to their knowledge of their treatment. This is not much of a concern in nonhuman studies.

In a **double-blind experiment**, the researchers administering the treatments and measuring the response are also unaware of which subjects are receiving which treatments. This prevents researchers who are interacting with the subjects from behaving differently toward them according to their treatments. Researchers sometimes have pet hypotheses, and they might treat experimental subjects in different ways depending on their hopes for the outcome. Moreover, many response variables are difficult to measure and require some subjective interpretation, which makes the results prone to a bias in favor of the researchers' wishes and expectations. Finally, researchers are naturally more interested in the treated subjects than the control subjects, and this increased attention can itself result in improved response. Reviews of medical studies have revealed that studies carried out without double-blinding exaggerated treatment effects by 16% on average, compared with studies carried out with double-blinding (Jüni et al. 2001).

Blinding is the process of concealing information from participants (sometimes including researchers) about which individuals receive which treatment.

Experiments on nonhuman subjects are also prone to bias from lack of blinding. Bebarta et al. (2003) reviewed 290 two-treatment experiments carried out on animals or on cell lines. They found that the odds of detecting a positive effect of treatment were more than threefold higher in studies without blinding than in studies with blinding.[3] Blinding can be incorporated into experiments on nonhuman subjects by using coded tags that identify the subject to a "blind" observer without revealing the treatment (and then the observer measures units from different treatments in random order).

14.4 How to reduce the influence of sampling error

Assuming we have designed our experiment to minimize sources of bias, there is still the problem of detecting any treatment effects against the background of variation between individuals ("noise") caused by other variables. Such variability creates

3. This result probably overestimates the effects of a lack of blinding, because the experiments without blinding also tended to have confounding problems, such as a lack of randomization (Bebarta et al. 2003).

sampling error in the estimates, reducing precision and power. How can the effects of sampling error be minimized?

One way to reduce noise is to make the experimental conditions constant. Fix the temperature, humidity, and other environmental conditions, for example, and use only participants who are of the same age, sex, genotype, and so on. In field experiments, however, highly constant experimental conditions might not be feasible. Constant conditions might not be desirable, either. By limiting the conditions of an experiment, we also limit the generality of the results—that is, the conclusions might apply only under the conditions tested and not more broadly. Until recently, a significant source of bias in medical practice stemmed from the fact that many clinical tests of the effects of medical treatments were carried out only on men, yet the treatments were subsequently applied to women as well (e.g., McCarthy 1994).

In this section, we review replication, balance, and blocking, the three main statistical design procedures used to minimize the effects of sampling error. We also review a strategy to reduce the effect of noise by using extreme treatments.

Replication

Because of variability, **replication**—the repetition of every treatment on multiple experimental units—is essential. Without replication, we would not know whether response differences were due to the treatments or just chance differences between the treatments caused by other factors. Studies that use more units (i.e., that have larger sample sizes) will have smaller standard errors and a higher probability of getting the correct answer from a hypothesis test. Larger samples give more information, and more information gives better estimates and more powerful tests.

Replication is the application of every treatment to multiple, independent experimental units.

Replication is not just about the number of plants or animals used. True replication depends on the number of *independent* units in the experiment. An "experimental unit" is the independent unit to which treatments are assigned. Figure 14.4-1 shows three hypothetical experiments designed to compare the effects of two fertilizer treatments on plant growth. The lack of replication is obvious in the first design (top panel), because there is only one plant per treatment. You won't see many published experiments like it.

The lack of replication is less obvious in the second design (the middle panel of Figure 14.4-1). Although there are multiple plants per treatment in the second design, all plants in one treatment are confined to one chamber and all plants in the second treatment are confined to another chamber. If there are environmental differences between chambers (e.g., differences in light conditions or humidity) beyond those stemming from the treatment itself, then plants in the same chamber will be more similar in their responses than plants in different chambers, apart from any treatment effects. The plants in the same chamber are not independent. As a result, the chamber,

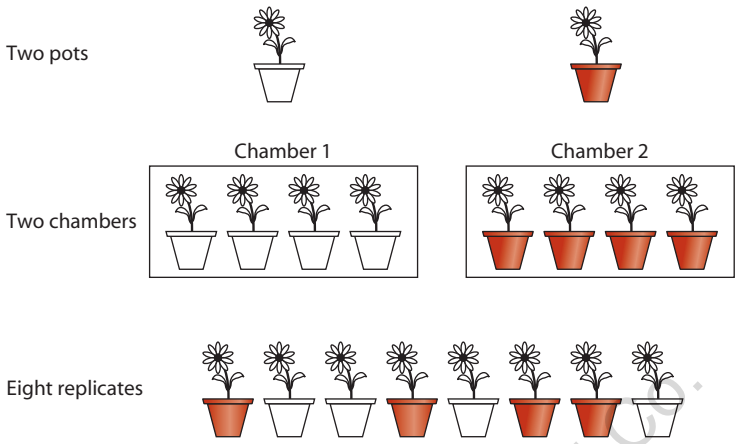

FIGURE 14.4-1 Three experimental designs used to compare plant growth under two fertilizer treatments (indicated by the shading of the pots). The upper ("two pots") and middle ("two chambers") designs are unreplicated.

not the plant, is the experimental unit in a test of fertilizer effects. Because there are only two chambers, one per treatment, the experiment is unreplicated.

Only the third design (the bottom panel) in Figure 14.4-1 is properly replicated, because here treatments have been randomly assigned to individual plants. A giveaway indicator of replication in the third design is *interspersion* of experimental units assigned different treatments, which is an expected outcome of randomization. Such interspersion is lacking in the two-chamber design (the middle panel in Figure 14.4-1), which is a clear sign of a replication problem.

An experimental unit might be a single animal or plant if individuals are randomly sampled and assigned treatments independently. Or, an experimental unit might be made up of a batch of individual organisms treated as a group, such as a field plot containing multiple individuals, a cage of animals, a household, a petri dish, or a family. Multiple individual organisms belonging to the same unit (e.g., plants in the same plot, bacteria in the same dish, members of the same family, and so on) should be considered together as a single replicate if they are likely to be more similar on average to each other than to individuals in separate units (apart from the effects of treatment).

Correctly identifying replicates in an experiment is crucial to planning its design and analyzing the results. Erroneously treating the single organism as the independent replicate when the chamber (Figure 14.4-1) or field plot is the experimental unit will lead to calculations of standard errors and P-values that are too small. This is pseudoreplication, as discussed in Interleaf 2.

From the standpoint of reducing sampling error, more replication is always better. As proof, examine the formula for the standard error of the difference between two sample mean responses to two treatments, $\overline{Y}_1 - \overline{Y}_2$:

$$\mathrm{SE}_{\overline{Y}_1-\overline{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

The symbols n_1 and n_2 refer to the number of experimental units, or replicates, in each of the two treatments. Based on this equation, increasing n_1 and n_2 directly reduces the standard error, increasing precision. Increased precision yields narrower confidence intervals and more powerful tests of the difference between means. On the other hand, increasing sample size also has costs in terms of time, money, and even lives. We discuss how to plan a sufficient sample size in more detail in Section 14.7.

Balance

A study design is **balanced** if all treatments have the same sample size. Conversely, a design is *un*balanced if there are *un*equal sample sizes between treatments.

In a *balanced* experimental design, all treatments have equal sample size.

Balance is a second way to reduce the influence of sampling error on estimation and hypothesis testing. To appreciate this, look again at the equation for the standard error of the difference between two treatment means (given on page 433). For a fixed total number of experimental units, $n_1 + n_2$, the standard error is smallest when the quantity

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

is smallest, which occurs when n_1 and n_2 are equal. Convince yourself that this is true by plugging in some numbers. For example, if the total number of units is 20, the quantity $1/n_1 + 1/n_2$ is 0.2 when $n_1 = n_2 = 10$, but it is 1.05 when $n_1 = 19$ and $n_2 = 1$. With better balance, the standard error is much smaller.

To estimate the difference between two groups, we need precise estimates of the means of *both* groups. With an unbalanced design, we may know the mean of one group with great precision, but this does not help us much if we have very little information about the other group that we're comparing it with. Balance allocates the sampling effort in the optimal way.

Nevertheless, the precision of an estimate of a difference between groups always increases with larger sample sizes, even if the sample size is increased in only one of two groups. But for a fixed total number of subjects, the optimal allocation is to have an equal number in each group.

Balance has other benefits, which we discuss elsewhere in the book. For example, the methods based on the normal distribution for comparing population means are most robust to departures from the assumption of equal variances when designs are balanced or nearly so (see Chapters 12 and 15).

Blocking

Blocking is an experimental design strategy used to account for extraneous variation by dividing the experimental units into groups, called **blocks** or strata, that

share common features. Within blocks, treatments are assigned randomly to experimental units. Blocking essentially repeats the same completely randomized experiment multiple times, once for each block, as shown schematically in Figure 14.4-2. Differences between treatments are evaluated only within blocks. In this way, much of the variation arising from differences between blocks is accounted for and won't reduce the power of the study.

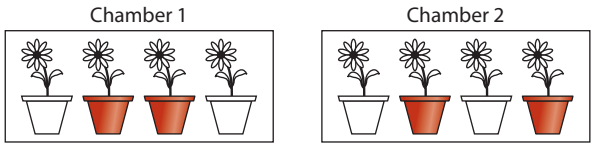

FIGURE 14.4-2 An experimental design incorporating blocking to test effects of fertilizer on plant growth (see Figure 14.4-1). Shading of the pots indicates which fertilizer treatment each plant received. Chambers might differ in unknown ways and add unwanted noise to the experiment. To remove the effects of such variation, carry out the same completely randomized experiment separately within each chamber. In this design, each chamber represents one block.

The women participating in the nonoxynol-9 HIV study discussed in Example 14.2 were grouped according to the clinic they attended. This made sense because there were age differences between women attending different clinics as well as differences in condom usage and sexual practices, all of which are likely to affect HIV transmission rates (van Damme et al. 2002). Blocking removes the variation in response among clinics, allowing more precise estimates and more powerful tests of the treatment effects.

Blocking is the grouping of experimental units that have similar properties. Within each block, treatments are randomly assigned to experimental units.

The *paired* design for two treatments (Chapter 12) is an example of blocking. In a paired design, both of two treatments are applied to each plot or other experimental unit representing a block. The difference between the two responses made on each block is the measure of the treatment effect.

The **randomized block design** is analogous to the paired design, but it can have more than two treatments, as shown in Example 14.4A.

The *randomized block design* is like a paired design but for more than two treatments.

EXAMPLE 14.4A Holey waters

The compact size of water-filled tree holes, which can harbor diverse communities of aquatic insect larvae, makes them useful microcosms for ecological experiments. Srivastava and Lawton (1998) made artificial tree holes from plastic that mimicked the buttress tree holes of European beech trees (see image on right). They placed the plastic holes next to trees in a forest in southern England to examine how the amount of decaying leaf litter present in the holes affected the number of insect eggs deposited (mainly by mosquitoes and hover flies) and the survival of the larvae emerging from those eggs. Leaf litter is the source of all nutrients in these holes, so increasing the amount of litter might result in more food for the insect larvae. There were three different treatments. In one treatment (LL), a low amount of leaf litter was provided. In a second treatment (HH), a high level of debris was provided. In the third treatment (LH), leaf litter amounts were initially low but were then made high after eggs had been deposited. A randomized block design was used in which artificial tree holes were laid out in triplets (blocks). Each block consisted of one LL tree hole, one HH tree hole, and one LH tree hole. The location of each treatment within a block was randomized, as shown in Figure 14.4-3.

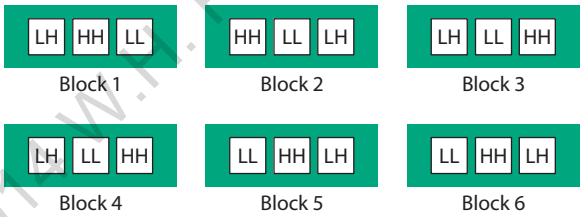

FIGURE 14.4-3 Schematic of the randomized block design used in the tree-hole study of Example 14.4A. Each block of three tree holes was placed next to its own beech tree in the woods. Within blocks, the three treatments were randomly assigned to tree holes.

As in the paired design, treatment effects in a randomized block design are measured by differences between treatments exclusively within blocks, a strategy that minimizes the influence of variation among blocks.

In the randomized block design, each treatment is applied once to every block. By accounting for some sources of sampling variation, such as the variation among trees, blocking can make differences between treatments stand out. In Chapter 18, we discuss in greater detail how to analyze data from a randomized block design.

Blocking is worthwhile if units within blocks are relatively homogeneous, apart from treatment effects, and units belonging to different blocks vary because of envi-

ronmental or other differences. For example, blocks can be made up of any of these units:

- Field plots experiencing similar local environmental conditions
- Animals from the same litter
- Aquaria located on the same side of the room
- Patients attending the same clinic
- Runs of an experiment executed on the same day

One potential drawback to blocking might occur if the effects of one treatment contaminate the effects of the other in the same block. For example, watering one half of a block might raise the soil humidity of the adjacent, unwatered half. Experiments should be designed carefully to minimize contamination.

Extreme treatments

Treatment effects are easiest to detect when they are large. Small differences between treatments are difficult to detect and require larger samples, whereas larger treatment differences are more likely to stand out against random variability within treatments. Therefore, one strategy to enhance the probability of detecting differences in an experiment is to include extreme treatments. Example 14.4B shows why this might be.

EXAMPLE 14.4B Plastic hormones

Bisphenol-A, or BPA, is an estrogenic compound found in plastics widely used to line food and drink containers and in dental sealants. Human daily exposures are typically in the range of 0.5–1 μg/kg body weight (Gray et al. 2004). Sakaue et al. (2001) measured sperm production of 13-week-old male rats exposed to fixed daily doses of BPA between 0 and 2000 μg/kg body weight for six days. The results are shown in a dose–response curve in Figure 14.4-4.

This experiment included doses much higher than the typical doses faced by humans at risk, a strategy that enhanced the ability to detect an effect of BPA. For example, Figure 14.4-4 shows that there was a much larger difference in mean sperm production between the 0 and 2000 μg/kg groups than between the 0 and 0.002 μg/kg treatments. If the experimenter were to design a study to compare just one of these doses with the control, using 200 or 2000 μg/kg would yield the most power, because they show the largest difference in sperm production from the control.

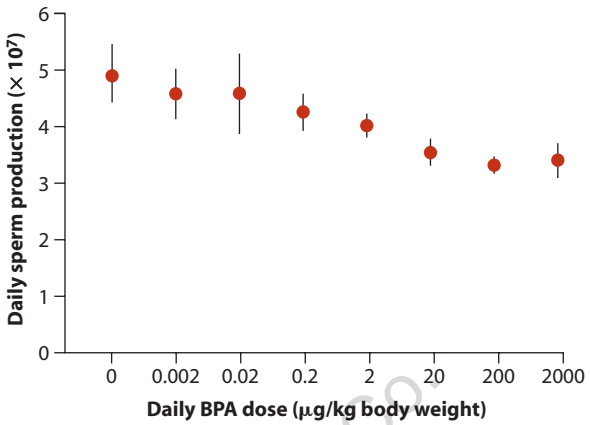

FIGURE 14.4-4
A dose-response curve showing the results of an experiment measuring the rates of sperm production of male rats exposed to fixed daily doses of bisphenol-A (BPA) (Sakaue et al. 2001). Symbols are the mean ±1 SE.

A larger dose, or stronger treatment, can increase the probability of detecting a response. But be aware that the effects of a treatment do not always scale linearly with the magnitude of a treatment. The effects of a large dose may be qualitatively different from those of a smaller, more realistic dose. Still, as a first step, extreme treatments can be a very good way to detect whether one variable has any effect at all on another variable.

14.5 Experiments with more than one factor

Up to now, we have considered only experiments that focus on measuring and testing the effects of a single factor. A **factor** is a single treatment variable whose effects are of interest to the researcher. However, many experiments in biology investigate more than one factor, because answering two questions from a single experiment rather than just one makes more efficient use of time, supplies, and other costs.

Another reason to consider experiments with multiple factors is that the factors might interact. When operating together, the factors might have synergistic or inhibitory effects not seen when each factor is tested alone. For example, human activity has driven global increases in atmospheric CO_2 and temperature, as well as greater nitrogen deposition and precipitation. Increases in all of these factors have been shown to stimulate plant growth by experimental studies in which each treatment variable was examined separately. But what are the effects of these factors in combination? The only way to answer this is to design experiments in which more than one factor is manipulated simultaneously. If the climate variables interact when influencing plant growth, then their joint effects can be very different from their separate effects (Shaw et al. 2002).

A *factor* is a single treatment variable whose effects are of interest to the researcher.

The factorial design is the most common experimental design used to investigate more than one treatment variable, or factor, at the same time. In a **factorial design**, every combination of treatments from two (or more) treatment variables is investigated.

An experiment having a *factorial design* investigates all treatment combinations of two or more variables. A factorial design can measure interactions between treatment variables.

The main purpose of a factorial design is to evaluate possible interactions between variables. An **interaction** between two explanatory variables means that the effect of one variable on the response depends on the state of a second variable. Example 14.5 illustrates an interaction in a factorial design.

EXAMPLE 14.5 Lethal combination

Frog populations are declining everywhere, spawning research to identify the causes. Relyea (2003) looked at how a moderate dose (1.6 mg/l) of a commonly used pesticide, carbaryl (Sevin), affected bullfrog tadpole survival. In particular, the experiment asked how the effect of carbaryl depended on whether a native predator, the red-spotted newt, was also present. The newt was caged and could cause no direct harm, but it emitted visual and chemical cues that are known to affect tadpoles. The experiment was carried out in 10-liter tubs, each containing 10 tadpoles. The four combinations of pesticide treatment (carbaryl vs. water only) and predator treatment (present or absent) were randomly assigned to tubs. For each combination of treatments, there were four replicate tubs. The effects on tadpole survival are displayed in Figure 14.5-1.

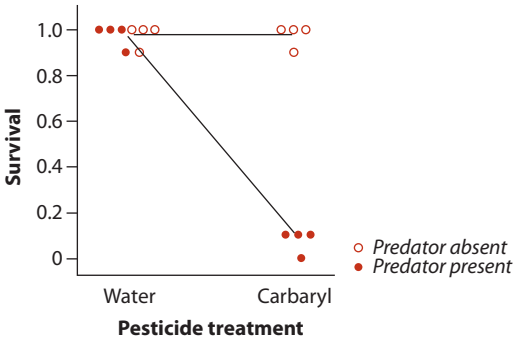

FIGURE 14.5-1
Interaction between the effects of the pesticide (carbaryl) and predator (red-spotted newt) treatments on tadpole survival. Each point gives the fraction of tadpoles in a tub that survived. Lines connect mean survival in the two pesticide treatments, separately for each predator treatment.

The tub, not the individual tadpole, is the experimental unit, because tadpoles sharing the same tub are not independent. The results showed that survival was high, except when pesticide was applied together with the predator—neither treatment

alone had much effect (Figure 14.5-1). Thus, the two treatments, predation and pesticide, seem to have interacted—that is, the effect of one variable depends on the state of the other variable. An experiment investigating the effects of the pesticide only would have measured little effect at this dose. Similarly, an experiment investigating the effect of the predator only would not have seen an effect on survival.

An *interaction* between two (or more) explanatory variables means that the effect of one variable depends upon the state of the other variable.

A factorial design can still be worthwhile even if there is no interaction between explanatory variables. In this case, there are efficiency advantages because the same experimental units can be used to measure the effect of two (or more) variables simultaneously.

14.6 What if you can't do experiments?

Experimental studies are not always feasible, in which case we must fall back upon observational studies. Observational studies can be very important, because they detect patterns and can help generate hypotheses. The best observational studies incorporate all of the features of experimental design used to minimize bias (e.g., simultaneous controls and blinding) and the impact of sampling error (e.g., replication, balance, blocking, and even extreme treatments), except for one: randomization. Randomization is out of the question because, in an observational study, the researcher does not assign treatments to subjects. Instead, the subjects come as they are.

Match and adjust

Without randomization, minimizing bias resulting from confounding variables is the greatest challenge of observational studies. Two types of strategies are used to limit the effects of confounding variables on a difference between treatments in a controlled observational study. One strategy, commonly used in epidemiological studies, is **matching**. With matching, every individual in the target group with a disease or other health condition is paired with a corresponding healthy individual who has the same measurements for known confounding variables, such as age, weight, sex, and ethnic background (Bland and Altman 1994).

Matching is often used when designing case-control studies. Recall from Chapter 9 that in a case-control study, exposure to one or more possible causal factors is compared between a sample of individuals having a disease (the cases) and a second sample of participants not having the disease (the controls). Matching ensures that the cases and controls are otherwise similar. For example, Dziekan et al. (2000) investigated possible causes of a hospital outbreak of antibiotic-resistant *Staphylococcus.*

The 67 infected cases were each paired with a control individual matched for age, sex, hospital admission date, and admission department.

Matching reduces bias by limiting the contribution of suspected confounding variables to differences between treatments. Unlike randomization in an experiment, matching in an observational study does not account for all confounding variables, only those explicitly used to match participants. Thus, while matching reduces bias, it does not eliminate bias. Matching also reduces sampling error by grouping experimental units into similar pairs, analogous to blocking in experimental studies. It is with such a matched case-control design that the link between smoking and lung cancer was convincingly demonstrated.

With *matching*, every individual in the treatment group is paired with a control individual having the same or closely similar values for the suspected confounding variables.

In a weaker version of this approach, a comparison group is chosen that has a frequency distribution of measurements for each confounding variable that is similar to that of the treatment group, but no pairing takes place. For example, attention deficit/hyperactivity disorder (ADHD) is often treated with stimulants, such as amphetamines. Biederman et al. (2009) carried out an observational study to examine the psychiatric impacts later in life of stimulant treatment. A sample of ADHD youths who had been treated with stimulants was compared with a control sample of untreated ADHD individuals that was similar to the treated group in the distribution of ages, sex, ethnic background, sensorimotor function, other psychiatric conditions, and IQ.

The second strategy used to limit the effects of confounding variables in a controlled observational study is adjustment, in which statistical methods such as analysis of covariance (Chapter 18) are used to correct for differences between treatment and control groups in suspected confounding variables. For example, LaCroix et al. (1996) compared the incidence of cardiovascular disease between two groups of older adults: those who walked more than four hours per week and those who walked less than one hour per week. The ages of the adults were not identical in the two groups, and this could affect the results. To compensate, the authors examined the relationship between cardiovascular disease and age within each exercise group, so that they could compare the predicted disease rates in the two groups for adults of the same age. This approach is discussed in more detail in Chapter 18.

14.7 Choosing a sample size

A key part of planning an experiment or observational study is to decide how many independent units or participants to include. There is no point in conducting a study whose sample size is too small to detect the expected treatment effect. Equally, there

is no point in making an estimate if the confidence interval for the treatment effect is expected to be extremely broad because of a small sample size. Using too many participants is also undesirable, because each replicate costs time and money, and adding one more might put yet another individual in harm's way, depending on the study. If the treatment is unsafe, as the spermicide nonoxynol-9 appears to be (Example 14.2), then we want to injure as few people or animals as possible in coming to this conclusion. Ethics boards and animal-care committees require researchers to justify the sample sizes for proposed experiments. How is the decision made? Here in Section 14.7, we answer this question for two objectives: when the goal is to achieve a predetermined level of *precision* of an estimate of treatment effect, or when we want to achieve predetermined *power* in a test of the null hypothesis of no treatment effect. We focus here on techniques for studies that compare the means of two groups. Formulas to help plan experiments for some other kinds of data are given in the Quick Formula Summary (Section 14.9).

An important part of planning an experiment or observational study is choosing a sample size that will give sufficient power or precision.

Plan for precision

A frequent goal of studies in biology is to estimate the magnitude of the treatment effect as precisely as possible. Planning for precision involves choosing a sample size that yields a confidence interval of expected width. Typically, we hope to set the bounds as narrowly as we can afford.

By way of example, let's develop a plan for a two-treatment comparison of means. Let the unknown population mean of the response variable be μ_1 in the treatment group of interest and μ_2 in the control group. When the results are in, we will compute the sample means $\overline{Y}_1$ and $\overline{Y}_2$ and use them to calculate a 95% confidence interval for $\mu_1 - \mu_2$, the difference between the population means of the treatment and control groups. To simplify matters somewhat, we will assume that the sample sizes in both treatments are the same number, n. Let's also assume that the measurement in the two populations is normally distributed and has the same standard deviation, σ.

In this case, a 95% confidence interval for $\mu_1 - \mu_2$ will take the form

$$\overline{Y}_1 - \overline{Y}_2 \pm \text{ margin of error,}$$

where "margin of error" is half the width of the confidence interval. Planning for precision involves deciding in advance how much uncertainty we can tolerate. Once we've decided that, then the sample size needed in each group is approximately

$$n = 8\left(\frac{\sigma}{\text{margin of error}}\right)^2.$$

This formula is derived from the 2SE rule of thumb that was introduced in Section 4.3.[4] According to this formula, a larger sample size is needed if σ, the standard deviation within groups, is large than if it is small. Additionally, a larger sample size is needed to achieve a high precision (a narrow confidence interval) than to achieve a low precision.

A major challenge in planning sample size is that key factors, like σ, are not known. Typically, a researcher makes an educated guess for these unknown parameters based on pilot studies or previous investigations. (If no information is available, then consider carrying out a small pilot study first, before attempting a large experiment.)

For example, let's plan an experiment to measure the effect of diet on the eye span of male stalk-eyed flies (Example 11.2). The planned experiment will randomly place individual fly larvae into cups containing either corn or spinach. The target parameter is the difference between mean eye spans in the two diet treatments, $\mu_1 - \mu_2$. Assume that we would like to obtain a 95% confidence interval for this difference whose expected margin of error is 0.1 mm (i.e., the desired full width of the confidence interval is 0.2 mm). How many male flies should be used in each treatment to achieve this goal?

Our sample estimate for σ was about 0.4, based on the sample of nine individuals in Example 11.2. Using these values gives

$$n = 8\left(\frac{\sigma}{\text{margin of error}}\right)^2 = 8\left(\frac{0.4}{0.1}\right)^2 = 128.$$

This is the sample size in each treatment, so the total number of male flies would be 256. At this point, we would need to decide whether this sample size is feasible in an experiment. If not, then there might be no point in carrying out the experiment. Alternatively, we could revisit the desired width of the 95% confidence interval. That is, could we be satisfied with a higher margin of error? If so, then we should decide on this new width and then recalculate n.

After all this planning, imagine that the experiment is run and we now have our data. Will the confidence interval we calculate have the precision we planned for? There are two reasons that it probably won't. First, 0.4 was just an educated guess for the value of σ to help our planning, and it was based on only nine individuals. The true value of σ in the population might be larger or smaller. Second, even if we were lucky and the true value of σ really is close to 0.4, the within-treatment standard deviation s from the experiment will not equal 0.4 because of sampling error. The resulting confidence interval will be narrower or wider accordingly. The probability that the width of the resulting confidence interval is less than or equal to the desired width is only about 0.5. To increase the probability of obtaining a confidence interval no wider than the desired interval width, we would need an even larger sample size.

Figure 14.7-1 shows the general relationship between the expected precision of the 95% confidence interval and n, the sample size in each of two groups. The

4. The margin of error is approximately twice the standard error of the difference between sample means (2SE), or $2\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{n})} = 2\sqrt{2\sigma^2/n} = \sqrt{8\sigma^2/n}$. Solving for n gives the rule in the text.

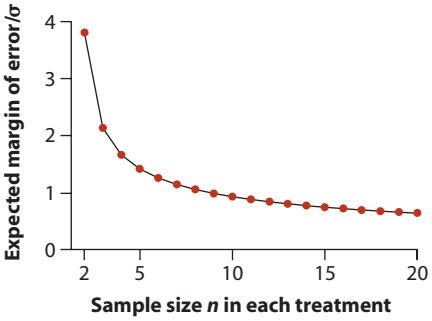

FIGURE 14.7-1
Expected precision of the 95% confidence interval for the difference between two treatment means depending on sample size n in each treatment. The vertical axis is given in standardized units, (margin of error)/σ. We calculated the expected confidence interval using the t-distribution, rather than with the 2SE approximation.

variable on the vertical axis is standardized as margin of error divided by σ. The effect of sample size from $n = 2$ to $n = 20$ is shown.

The graph shows that very small sample sizes lead to very wide interval estimates of the difference between treatment means. More data gives better precision. Note also that interval precision initially declines rapidly with increasing sample size (e.g., from $n = 2$ to $n = 10$), but it then declines more slowly (e.g., from $n = 10$ to $n = 20$). Precision is 0.63 at $n = 20$, but it drops to 0.40 by $n = 50$, to 0.28 by $n = 100$, and to 0.20 by $n = 200$. Thus, we get diminishing returns by increasing the sample size past a certain point.

Plan for power

Next we consider choosing a sample size based on a desired probability of rejecting a false null hypothesis—that is, planning a sample based on a desired power. Imagine, for example, that we want to test the following hypotheses on the effect of diet on eye span in stalk-eyed flies.

H_0: $\mu_1 - \mu_2 = 0$.
H_A: $\mu_1 - \mu_2 \neq 0$.

The null hypothesis is that diet has no effect on mean eye span. The power of this test is the probability of rejecting H_0 if it is false. Planning for power involves choosing a sample size that would have a high probability of rejecting H_0 if the absolute value of the difference between the means, $|\mu_1 - \mu_2|$, is at least as great as a specified value D. The value for D won't be the true difference between the means; it is just the minimum we care about. By specifying a value for D in a sample size calculation, we are deciding that we aren't much interested in rejecting the null hypothesis of no difference if $|\mu_1 - \mu_2|$ is smaller than D.

A conventional power to aim for is 0.80. That is, if H_0 is false, we aim to demonstrate that it is false in 80% of the experiments (the other 20% of experiments would fail to reject H_0 even though it is false). If we aim for a power of 0.80 and a conventional significance level of $\alpha = 0.05$, then a quick approximation to the planned sample size n in each of two groups is

$$n \approx 16\left(\frac{\sigma}{D}\right)^2$$

(Lehr 1992). This formula assumes that the two populations are normally distributed and have the same standard deviation (σ), which we are forced to assume is known. A more exact formula is provided in the Quick Formula Summary (Section 14.9), which also allows you to choose other values for power and significance level.

For a given power and significance level, a larger sample size is needed when the standard deviation σ within groups is large, or if the minimum difference that we wish to detect is small.

Let's return to our experiment to test the effect of diet on the eye span of male stalk-eyed flies. We would like to reject H_0 at $\alpha = 0.05$ with probability 0.80 if the absolute value of the difference between means were truly $D = |\mu_1 - \mu_2| = 0.2$ mm. How many males should be used in each treatment?

Let's assume again that $\sigma = 0.4$. Using this value in the equation for power gives

$$n = 16\left(\frac{0.4}{0.2}\right)^2 = 64.$$

This is the number in each treatment, so the total number of males needed in the experiment would be 128.

These power calculations assume that we know the standard deviation (σ), which is stretching the truth. For this and other reasons, we must always view the results of power calculations with a great deal of caution. The calculations provide useful guidelines, but they do not give infallible answers.

We have explored only the sample sizes needed to compare the means of two groups, but similar methods are available for other kinds of statistical comparisons as well. Sample sizes for desired precision and power are available for one- and two-sample means, proportions, and odds ratios in the Quick Formula Summary (Section 14.9). A variety of computer programs are available to calculate sample sizes when planning for power and precision. A good place to start investigating these programs is http://www.divms.uiowa.edu/~rlenth/Power/.

Plan for data loss

The methods given here in Section 14.7 for planning sample sizes refer to sample sizes still available at the *end* of the experiment. But some experimental individuals may die, leave the study, or be lost between the start and the end of the study. The starting sample sizes should be made even larger to compensate.

14.8 Summary

- In an experimental study, the researcher assigns treatments to subjects.
- The purpose of an experimental study is to examine the causal relationship between an explanatory variable, such as treatment, and a response variable. The virtue of experiments is that the effect of treatment can be isolated by randomizing the effects of confounding variables.

- A confounding variable masks or distorts the causal relationship between an explanatory variable and a response variable in a study.
- A clinical trial is an experimental study involving human participants.
- Experiments should be designed to minimize bias and limit the effects of sampling error.
- Bias in experimental studies is reduced by the use of controls, by randomizing the assignment of treatments to experimental units, and by blinding.
- In a completely randomized experiment, treatments are assigned to experimental units by randomization. Randomization reduces the bias caused by confounding variables by making nonexperimental variables equal (on average) between treatments.
- The effect of sampling error in experimental studies is reduced by replication, by blocking, and by balanced designs.
- A randomized block design is like a paired design but for more than two treatments.
- The use of extreme treatments can increase the power of the experiment to detect a treatment effect.
- Observational studies should employ as many of the strategies of experimental studies as possible to minimize bias and limit the effect of sampling error.
- Although randomization is not possible in observational studies, the effects of confounding variables can be reduced by matching and by adjusting for differences between treatments in known confounding variables.
- A factorial design is used to investigate the interaction between two or more treatment variables. The factorial design includes all possible combinations of the treatment variables.
- When planning an experiment, the number of experimental units to include can be chosen so as to achieve the desired width of confidence interval for the difference between treatment means.
- Alternatively, the number of experimental units to include when planning an experiment can be chosen so that the probability of rejecting a false H_0 (power) is high for a specified magnitude of the difference between treatment means.
- Compensate for possible data loss when planning sample sizes for an experiment.

14.9 Quick Formula Summary

Planning for precision

Planned sample size for a 95% confidence interval of a proportion

What is it for? To set the sample size of a planned experiment to achieve approximately a specified half-width ("margin of error") of a 95% confidence interval for a proportion p.

What does it assume? The population proportion p is not close to zero or one, and n is large.

Formula: $n \approx \dfrac{4p(1-p)}{(\text{margin of error})^2}$, where p is the proportion being estimated and "margin of error" is the half-width of the confidence interval for the proportion p. For the most conservative scenario, set $p = 0.50$ when calculating n. The symbol $\approx$ stands for "is approximately equal to."

Planned sample size for a 95% confidence interval of a log-odds ratio

What is it for? To set the sample size n in each of two groups for a planned experiment to achieve approximately a specified half-width ("margin of error") of a 95% confidence interval for a log-odds ratio, $\ln(OR)$.

What does it assume? Sample size n is the same in both groups.

Formula: $n \approx \dfrac{4}{(\text{margin of error})^2}\left(\dfrac{1}{p_1} + \dfrac{1}{1-p_1} + \dfrac{1}{p_2} + \dfrac{1}{1-p_2}\right)$, where "margin of error" is the half-width of the confidence interval for $\ln(OR)$, and p_1 and p_2 are the probabilities of success in the two treatment groups.

Planned sample size for a 95% confidence interval of the difference between two proportions

What is it for? To set the sample size n in each of two groups for a planned experiment to achieve approximately a specified half-width ("margin of error") of a 95% confidence interval for a difference between two proportions, $p_1 - p_2$. This is an alternative approach to the one that uses a log-odds ratio to compare the proportion of successes in two treatment groups.

What does it assume? Sample size n is the same in both groups.

Formula: $n \approx \dfrac{8\bar{p}(1-\bar{p})}{(\text{margin of error})^2}$, where "margin of error" is the half-width of the confidence interval for $p_1 - p_2$. p_1 and p_2 are the probabilities of success in the two treatment groups, and $\bar{p}$ is the average of the two proportions—that is, $\bar{p} = (p_1 + p_2)/2$.

Planned sample size for a 95% confidence interval of the mean

What is it for? To set the sample size of a planned experiment to achieve approximately a specified half-width ("margin of error") of a 95% confidence interval for a mean μ.

What does it assume? The population is normally distributed with known standard deviation σ.

Formula: $n \approx 4\left(\dfrac{\sigma}{\text{margin of error}}\right)^2$, where n is the planned sample size, and "margin of error" is the half-width of the confidence interval for the mean μ.

Planned sample size for a 95% confidence interval of the difference between two means

What is it for? To set the sample size of a planned experiment so as to achieve approximately a specified half-width ("margin of error") of the 95% confidence interval for $\mu_1 - \mu_2$.

What does it assume? Populations are normally distributed with equal standard deviations σ. The value of σ is known. Sample size n is the same in both groups.

Formula: $n \approx 8\left(\dfrac{\sigma}{\text{margin of error}}\right)^2$, where n is the planned sample size within each group, and "margin of error" is the half-width of the confidence interval for the difference between means.

Planning for power

Planned sample size for a binomial test of 80% power at $\alpha = 0.05$

What is it for? To set the sample size n of a planned experiment to achieve approximately a power of 0.80 in a binomial test at $\alpha = 0.05$.

What does it assume? The proportion p_0 under the null hypothesis is not close to zero or one, and n is not small. Sample size n is the same in both groups.

Formula: $n \approx \dfrac{8p_0(1 - p_0)}{D^2}$, where p_0 is the proportion under the null hypothesis, and $D = p - p_0$ is the predetermined difference we wish to be able to detect between the population parameter p and that specified under the null hypothesis.

Planned sample size for 2 × 2 contingency test of 80% power at $\alpha = 0.05$

What is it for? To set the sample size of a planned experiment so as to achieve approximately a power of 0.80 at $\alpha = 0.05$ in a contingency test of the difference between the proportion of successes in two treatment groups (or, equivalently, a test that the odds ratio equals one).

What does it assume? The average probability of success in the two treatment groups is known. Sample size n is the same in both treatment groups.

Formula: $n \approx \dfrac{8\bar{p}(1 - \bar{p})}{D^2}$, where $\bar{p}$ is the average of the two probabilities of success [i.e., $\bar{p} = (p_1 + p_2)/2$], and $D = p_1 - p_2$ is the predetermined difference we wish to be able to detect between the two proportions.

Planned sample size for a one-sample or paired *t*-test of 80% power at $\alpha = 0.05$

What is it for? To set the sample size of a planned experiment so as to achieve approximately a power of 0.80 in a one-sample or paired *t*-test at $\alpha = 0.05$.

What does it assume? The population is normally distributed with standard deviation σ. The value of σ is known.

Formula: $n \approx 8\left(\dfrac{\sigma}{D}\right)^2$, where n is the sample size within each group, and $D = \mu$ is the predetermined value of the mean (or the mean difference in the case of a paired test) that we wish to detect.

Planned sample size for a two-sample *t*-test of 80% power at $\alpha = 0.05$

What is it for? To set the sample size of a planned experiment so as to achieve approximately a power of 0.80 in a two-sample *t*-test at $\alpha = 0.05$.

What does it assume? Populations are normally distributed with equal standard deviation σ. Sample size n is the same in both groups.

Formula: $n \approx 16\left(\frac{\sigma}{D}\right)^2$, where n is the sample size within each group, and $D = |\mu_1 - \mu_2|$ is the predetermined difference between means we wish to detect.

PRACTICE PROBLEMS

1. Identify which goal of experimental design (i.e., reducing bias or limiting sampling error) is aided by the following procedures:
- **a.** Using a genetically uniform animal stock to test treatment effects
- **b.** Using a completely randomized design
- **c.** Grouping related experimental units together
- **d.** Taking the response measurements while unaware of the treatments assigned to experimental units
- **e.** Using a computer to randomly assign treatments to experimental units within each block

2. Using a coin toss for each unit, assign two hypothetical treatments to eight experimental units.
- **a.** Write the sequence of eight assignments you ended up with.
- **b.** Did you end up with an equal number of units in each treatment?
- **c.** What is the probability of an unbalanced design using this approach?
- **d.** Recommend a procedure for randomly assigning treatments to units that always results in a balanced design.

3. A series of plots were placed in a large agricultural field in preparation for an experiment to investigate the effects of three fertilizers differing in their chemical composition. Before assigning treatments, it was noticed that plots differed along a moisture gradient. What strategy would you suggest the researchers implement to minimize the impact of this gradient on the ability to measure a treatment effect? Explain with an illustration the experimental design you would recommend.

4. You read the following statement in a journal article: "On the basis of an alpha level of 0.05 and a power of 80%, the planned sample size was 129 subjects in each treatment group." State in plain language what this means.

5. Example 12.4 described a study in which salmon were introduced to 12 streams with and without brook trout to investigate the effect of brook trout on salmon survival. Is this an experimental study or an observational study? Explain the basis for your reasoning.

6. Identify the consequences (i.e., increase, decrease, or none) that the following procedures are likely to have on both bias and sampling error in an observational study.
- **a.** Matching sampling units between treatment and control
- **b.** Increasing sample size
- **c.** Ensuring that the frequency distribution of subject ages is the same in the two treatments
- **d.** Using a balanced design

7. In 1899, the *British Medical Journal* (page 933) reported the results of a medical procedure involving the subcutaneous infusion of a salt solution for the treatment of extremely severe pneumonia: "Dr. Clement Penrose has tried the effect of subcutaneous salt infusions as a last extremity in severe cases of pneumonia. He continues this treatment with inhalations of oxygen. He has had experience of three cases, all considered hopeless, and succeeded in saving one. In the other two the prolongation of life and the relief of symptoms were so marked that Dr. Penrose regretted that the treatment had not been employed earlier."
- **a.** Is this an experimental study? Why or why not?
- **b.** What design components might Dr. Penrose have included in an experiment to test the effectiveness of his treatment?

8. In a study of the effects of marijuana on the risk of cancer in oral squamous cells, Rosenblatt et al. (2004) examined 407 recent cases of the cancer from western Washington state. They also randomly sampled 615 healthy people from the same region having similar frequency distributions of age and sex as the cancer cases. They found that a similar proportion of the cancer cases (25.6%) and healthy participants (24.4%) reported ever having used marijuana (odds ratio = 0.9; 95% confidence interval, $0.6 < OR < 1.3$).

- **a.** What name is given to this type of study? Is it an experimental study or an observational study? Explain.
- **b.** Does this study include a control group? Explain.
- **c.** What was the purpose of ensuring that the healthy participants were similar in age and sex to the cancer cases?
- **d.** Can we conclude that marijuana does not cause cancer in oral squamous cells in this population?

9. After stinging its victim, the honeybee leaves behind the barbed stinger, poison sac, and muscles that continue to pump venom into the wound. Visscher et al. (1996) compared the effects of two methods of removing the stinger left behind: scraping off with a credit card or pinching off with thumb and index finger. A total of 40 stings were induced on volunteers. Twenty were removed with the credit card method, and 20 were removed with the pinching method. The size of the subsequent welt by each sting was measured after 10 minutes. All 40 measurements came from two volunteers (both authors of the study), each of whom received one treatment 10 times on one arm and the other treatment 10 times on the other arm. Pinching led to a slightly smaller average welt, but the difference between methods was not significant.

- **a.** All 40 measurements were combined to estimate means, standard errors, and the *P*-value for a two-sample *t*-test of the difference between treatment means. What is wrong with this approach?
- **b.** How should the data be analyzed? Describe how the quantities would be calculated and what type of statistical test would be used.
- **c.** Suggest two improvements to the experimental design.

10. What is the justification for including extreme doses well outside the range of exposures encountered by people at risk in a dose–response study on animals of the effects of a hazardous substance? What are the problems with this approach?

11. A strain of sweet corn has been genetically modified with a gene from the bacterium *Bacillus thuringiensis* (Bt) to express the protein Cry1Ab, which is toxic to caterpillars that eat the leaves. Unfortunately, the pollen of transformed corn plants contains the toxin, too. Corn pollen dusts the leaves of other plants growing nearby, where it might have negative effects on non-pest caterpillars. You are hired to conduct a study to measure the effects on monarch butterfly caterpillars of ingesting Bt-modified pollen that has landed on the leaves of milkweed, a plant commonly growing in or near cornfields. You decide to use a completely randomized design to compare the effect of two treatments on monarch pupal weight. In one treatment, you place potted milkweed plants in plots of Bt-modified corn, where their leaves receive pollen carrying the toxin. In the other treatment, you place milkweed plants in plots with ordinary corn that has not been transformed with the Bt gene. You place a monarch larva on every milkweed plant. Previous studies have estimated that the standard deviation of pupal weight in monarch butterflies is about 0.25 g.

a. Suppose your goal at the end of the experiment is to calculate a 95% confidence interval for the difference between treatments in mean monarch pupal weights. How many plots would you plan in each treatment if your goal was to produce a confidence interval for the difference in mean pupal weights between treatments having a total width of 0.4 g?

b. What sample size would you need if you decided that 0.4 was not precise enough, and that you wished to halve this interval to 0.2?

c. Imagine that your permits allow you to plant only five plots of Bt-transformed corn, so that the only way you can increase the total sample size for the whole experiment is to increase the number of plots in the ordinary corn treatment. To achieve the same width of confidence interval as in part (a), would the total sample size needed (both treatments combined) likely be greater, smaller, or no different from that calculated in part (a)? Explain.

d. In designing the experiment, why would you not simply place all the milkweed plants for one treatment at random locations in a single large Bt-transformed corn field, and all the milkweed plants for the other treatment at random locations in a single large normal corn field?

12. In the Bt and monarch study described in Practice Problem 11, how many plots would you plan per treatment if your goal were to carry out a test having 80% power to reject the null hypothesis of no treatment effects when the difference between treatments means is at least 0.25 g?

13. Consider the results of a six-year observational study that documented health changes related to homeopathic care (Spence and Thompson 2005). Homeopathic treatment was defined as "stimulating the body's autoregulatory mechanisms using microdoses of toxins." Every one of the 6544 patients in the study was assigned to a hospital outpatient unit for homeopathic treatment. Of these, 4627 patients (70.7%) reported positive health changes following treatment. Suggest a major improvement to the design of this study.

14. The fish species *Astyanax mexicanus* includes blind, cave-inhabiting populations whose eyes degenerate during embryonic development. To understand how eye degeneration worked, Yamamoto and Jeffery (2000) replaced the lens of the degenerate eye on one side (randomly chosen) of a blind cave fish embryo with a lens from the embryo of a "normal," sighted fish. This procedure was repeated on all individuals in a sample of blind cave fish. Final eye size was measured on both sides of each experimental fish, after embryonic development was complete. Remarkably, a normal-sized eye was restored on the transplant eyes of blind cave fish but not on the unmanipulated side. Based on the preceding description of a laboratory experiment, identify which of the six main strategies of experimental design (listed in Section 14.2) were incorporated.

15. Blaustein et al. (1997) used a field experiment to investigate whether increased UV-B radiation was a cause of amphibian deformities (see the photo at the beginning of this chapter). They measured long-toed salamanders either exposed to or shielded from natural UV-B radiation. It was not possible to carry out all replicates simultaneously, so the researchers carried them out over several days. They made sure that both treatments were included on each day. In their analysis, they grouped replicates together that were carried out on the same day.

a. By grouping experiments carried out on the same day, what experimental procedure were they using?

b. What is the main reason for adopting this procedure in an experimental study?

16. In 1976, Ewan Cameron and Linus Pauling (the only person to have won two unshared Nobel Prizes) published a paper showing that vitamin C was an effective treatment for some kinds of cancer. They measured the life spans of a sample of 100 patients who were given extra doses of vitamin C. As a control, they pulled the records of several hundred patients from the same clinic who had died from the same types of terminal

cancer, and who were matched to the vitamin C patients for their age, sex, and type of cancer. They found that the patients with extra vitamin C lived on average 2.7 times longer than the controls. A later study by Moertel et al. (1985) randomly assigned two treatments to cancer patients, supplemental vitamin C and control, and followed the patients with a double-blind study. This later study found no difference between the two groups for their life spans.

a. Give plausible reasons why the two studies might have found different results.

b. From the information given, which study is expected to give the most reliable results? Why?

ASSIGNMENT PROBLEMS

17. Identify the consequences (i.e., increase, decrease, or none) that the following procedures are likely to have on both bias and sampling error in an experimental study.

a. Assigning treatment to subjects alphabetically, not randomly

b. Increasing sample size

c. Calculating power

d. Applying every treatment to every experimental unit in random order

e. Using a sample of convenience instead of a random sample

f. Testing only one treatment group, without a control group

g. Using a balanced design

h. Informing the human participants which treatment they will receive

18. The experiment described in Example 12.2 compared antibody production in 13 male red-winged blackbirds before and after testosterone implants. The units of antibody levels were log 10^{-3} optical density per minute (ln[mOD/min]). The mean change in antibody production was $\bar{d} = 0.056$, and the standard deviation was $s_d = 0.159$. If you were assigned the task of repeating this experiment to test the hypothesis that testosterone changed antibody levels, what sample size (i.e., number of blackbirds) would you plan to ensure that a mean change of 0.05 units could be detected with probability 0.8? Explain the steps you took to determine this value.

19. Two clinical trials were designed to test the effectiveness of laser treatment for acne. Seaton et al. (2003) randomly divided participants into two groups. One group received the laser treatment, whereas the other group received a sham treatment. Orringer et al. (2004) used an alternate design in which laser treatment was applied to one side of the face, randomly chosen, and the sham treatment was applied to the other side. The number of facial lesions was the response variable.

a. Identify the main component of experimental design that differs between the two studies. Give the statistical term identifying this component in experimental design.

b. Under what circumstances would there be an advantage to using the "divided-face" design over the completely randomized (two-sample) design?

c. Assuming that the advantage identified in part (b) is met, can you think of a disadvantage of the divided-face design?[5]

20. Identify the consequences (i.e., increase, decrease, or none) that the following procedures are likely to have on both bias and sampling error in an observational study.

a. Planning for data loss

b. Taking measurements of the subjects while unaware of which subjects belong to which group

c. Including only one sex and age group in the study

5. Other than a possible social dilemma.

d. Adjusting for body size using analysis of covariance

21. Identify which goal of experimental design (i.e., reducing bias or limiting effects of sampling error) is aided by the following procedures:
a. Including extreme treatment levels
b. Using a paired design
c. Keeping room temperature constant in an experiment designed to test the effects of a pesticide on insect survival
d. Eliminating artifacts when designing the treatment of interest
e. Adding a sham operation group

22. Identify the particular feature that defines each of the following experimental designs, and list the specific advantages provided by the feature you identify.
a. Factorial design
b. Randomized block design
c. Completely randomized design

23. Kirsch (2010) argues that in double-blind clinical trials to test the effects of antidepressants, a large fraction of patients figure out whether they have been given the antidepressant or the placebo by noticing the presence or absence of known side effects of the antidepressant. Doctors evaluating the patients are also able to determine which treatment patients are receiving. How might this situation affect the results of the clinical trial? Specifically, is the treatment effect (difference between the means of the antidepressant and placebo treatments) likely to be overestimated, underestimated, or unaffected by this knowledge?

24. Michalsen et al. (2003) conducted a study to examine the effects of "leech therapy" for pain resulting from osteoarthritis of the knee. Two treatments were randomly assigned to 51 patients with osteoarthritis of the knee. Patients in the leech treatment received 4–6 medicinal leeches applied to the soft tissue of the affected knee in a single session. The animals were left to feed ad libitum until they detached themselves, on average 70 minutes later. Patients in the control treatment were given diclofenac gel and were told to apply it twice daily to the affected area. Pain was assessed by a questionnaire given by personnel unaware of the treatments applied to each patient. The results showed that seven days after the start of treatment, pain was significantly lower in the leech group.
a. Does this study include a control group? Explain.
b. Is this an experimental study or an observational study? Explain.
c. Is this a completely randomized design or a randomized block design? Explain.
d. Which strategy for reducing bias was not adopted in this study? How might its absence have affected the results?

25. Design a study to compare the reaction times of the left and right hands of right-handed people using a computer mouse. Two design choices are available to you. In the first, a sample of right-handed participants are randomly divided into two groups. Reaction time with the left hand is measured in one group, and reaction time with the right hand is tested in the other group.
a. What is the second design choice available to you?
b. Under what circumstances would the second choice be the preferred choice?
c. Assume that you decide to go with the completely randomized design and that, at the end of the experiment, you aim to calculate a 95% confidence interval for the difference between the mean reaction times of left and right hands. To achieve a confidence interval of specified width, what information would you require to plan an appropriate sample size?

26. Identify the single most significant flaw in each of the following experimental designs. Use statistical language to identify what's missing.
a. In a test of the effectiveness of acupuncture in treating migraine headaches, a random sample of patients at a migraine clinic were provided with a novel acupuncture treatment daily for six months. The patients were interviewed at the start of the study and at the end to determine whether there had been any change in the severity of their migraines.
b. In a modified study, a second sample of patients were chosen after the acupuncture

treatment was completed on the first set of patients. This second sample of patients received a placebo in pill form for six months. At the end of the study, perceived pain levels in the two groups were compared.

c. In a modified study, the sample of patients was divided into two groups according to gender. The women received the acupuncture treatment, and the men received the placebo medication in pill form. At the end of the experiment, perceived pain levels in the two groups were compared.

d. In a modified study, patients were randomly divided into two groups. One group received the acupuncture treatment, and the other received a fixed dose of placebo medication in pill form. At the end of the experiment, perceived pain levels in the two groups were compared.

27. Young et al. (2006) took measurements of subordinate female meerkats to determine the changes in reproductive physiology experienced by females that are evicted from their social groups. They compared evicted females and those not evicted in their level of plasma luteinizing hormone following a GnRH hormone challenge. They found that nine evicted females had a sample average of 6.2 mIU/ml (milli-International Units per milliliter) of plasma luteinizing hormone compared with 12.1 mIU/ml in 18 females that had not been evicted. The pooled sample variance was 28.4.

a. Is this an experimental study or an observational study? Explain.

b. The sample size was unequal between the two groups of females compared. How would this affect the power of a hypothesis test of the difference between group means compared with a more balanced design? Explain.

c. How would the imbalance of the sample sizes affect the width of the confidence interval for the difference between group means compared with a more balanced design? Explain.

d. If you were planning to repeat the comparison of plasma luteinizing hormone between these two groups of females, what sample size would you plan to achieve an expected half-width of 3 mIU/ml for a 95% confidence interval of the difference between means? Explain the steps you took to determine this value.

28. Diet restriction is known to extend life and reduce the occurrence of age-related diseases. To understand the mechanism better, you propose to carry out a study to look at the separate effects of age and diet restriction, and the interaction between age and diet restriction, on the activity of liver cells in rats. What experimental design should you consider employing? Why?

Data dredging

In the spoof journal *Annals of Improbable Research*, a satirical article reported on a study of the so-called butterfly effect (Inaudi et al. 1995). This effect, a mainstay of the popular representation of chaos theory, says that small initial causes, like the flapping of a butterfly's wings, can ultimately have large effects, like a hurricane, on the other side of the world. The fearless researchers set out to measure this effect by capturing several dozen butterflies and holding them in captivity in Switzerland. Each day, they checked the butterflies and recorded whether or not they flapped their wings. Then, using the lab's phone, they called their girlfriends in Paris each day to ask whether or not it was raining.[1] At the end of the study, the students tested each butterfly for an association between its daily flapping behavior and the daily weather in Paris. They found that the flapping days of one of the butterflies closely matched the rainfall days in Paris ($P < 0.05$). They exulted, "Not only have we proven that the butterfly effect exists, we have found the butterfly."

These guys were clearly joking, but statistically speaking, where did they go wrong? The answer is that they went "data dredging." They performed many statistical tests and eventually one of them was significant. Data dredging (also called "data snooping" or "data fishing") is the carrying out of many statistical tests in hope of finding at least one statistically significant result.

The problem with data dredging is that the probability of making *at least one* Type I error (i.e., of obtaining a false positive) is greater than the significance level α when many tests are conducted, if the null hypothesis is true (as it surely is in the butterfly example). Each hypothesis test has some chance of error, and these errors are compounded over multiple tests. There is a much larger probability of getting an error out of several tries than in any one try. By analogy, we might get away with playing Russian roulette once, but we would be unlikely to survive a month of playing once a night.

It's useful to do a few calculations to see how big the problem might be. The probability of making no Type I errors in N independent tests is $(1 - \alpha)^N$. Thus, the chance of making at least one Type I error from N independent tests is $1 - (1 - \alpha)^N$. This means that, if we use $\alpha = 0.05$ and carry out 20 independent tests of true null hypotheses, the probability that at least one of these tests will falsely reject the null hypothesis is about 65%. If we carry out 100

1. They continued the experiment "until the first phone bill reached our Office of Financial Services."

tests, then the chance of rejecting at least one of the null hypotheses becomes 99.4%, even if all the null hypotheses are true. With data dredging, a false positive result is almost inevitable.

Nevertheless, multiple testing is common in biology, and for good reasons. A dedicated experimentalist on human participants might measure many conceivable responses (e.g., blood pressure, body temperature, red blood cell count, white blood cell count, speed of recovery, appetite, and weight change) and perhaps even a few extra variables that might be long shots. The result is that the clinician might end up carrying out 10 or 20 tests of treatment effects, raising the probability of a false positive result. This level of multiple testing pales next to that seen in gene mapping. Locating a gene for a single trait, such as a genetic disease, typically involves thousands of statistical tests (one for each section of the genome). What should be done about the soaring Type I error rates resulting from so much testing?

The answer to this question depends on your goals. If your goal is simply to *explore* the data, to discover the possibilities but not to provide rigorous tests, then you need do nothing special about multiple testing except report the number of tests that you carried out and note which ones yielded a significant result. If you admit that you dredged the data, your results can still be useful. New hypotheses and unexpected discoveries can emerge from a thorough fishing expedition. However, the individual significant results that pop up from data dredging cannot yet be taken seriously, due to the high probability of one or more Type I errors. Some of the significant results might indeed be real, but it will be difficult to establish which ones. Rather, a new study must be carried out with new data to test any promising results that emerged from the exploratory approach. Another strategy sometimes used when exploring data is to divide the data randomly into two independent parts. One part is used for data dredging, and the other part is used to confirm any positive results suggested by the dredging.

If your goals from multiple testing are more rigorous (e.g., you want to determine which variable really did respond to treatment in a clinical trial, or which location in the genome really does contain a gene for a heritable disease), then steps must be taken to *correct* for the inflation of Type I error rates that occurs with multiple testing. The simplest way to accomplish this is to use a more stringent significance level—that is, one smaller than the usual $\alpha = 0.05$.

The most common correction for multiple comparisons is the **Bonferroni correction**. In the simplest version of this method, each test uses a significance level α^* rather than α, where

$$\alpha^* = \frac{\alpha}{\text{number of tests}}$$

For example, if we typically adopt the significance level $\alpha = 0.05$ when carrying out a single test, then to carry out 12 separate tests we should use the significance level $\alpha^* = 0.05/12 = 0.00417$ instead. In this case, we would reject H_0 in each test only if P were less than or equal to 0.00417. With the Bonferroni correction, the probability of getting at least one Type I error during the course of carrying out all 12 tests is approximately equal to the initial α-value (i.e., 0.05 in this case).

Keep in mind, though, that applying the Bonferroni correction greatly reduces the power of single tests. This is the price paid

for asking many questions of the data. More than ever, we should be mindful not to "accept the null hypothesis." It is okay to be skeptical when a null hypothesis is not rejected and power is so limited, but there is little to do about it except to repeat the study and look again.

Another, increasingly popular approach to correct for multiple comparisons is called the **false discovery rate** (**FDR**). To use this approach, carry out all of the multiple tests at a fixed significance level α (e.g., the usual 0.05). Gather all of the tests that yield a statistically significant result (i.e., all of the tests for which $P \leq \alpha$). We can call this subset of tests the "discoveries." The FDR estimates the proportion of discoveries that are false positives. In other words, the FDR is the proportion of tests for which the null hypothesis was rejected yet the null hypothesis was true. For example, Brem et al. (2005) carried out hundreds of statistical hypothesis tests of interactions between pairs of yeast genes. Of these tests, 225 yielded a statistically significant result (the "discoveries"). Using the false discovery rate method, they estimated that 12 of these 225 tests were false positives, leaving 213 "true" discoveries.

An extension of the FDR calculates a quantity called the q-value for each discovery. The q-value is analogous to a P-value, providing a measure of the strength of support from the data that the null hypothesis is false in a specific test. The smaller the q-value, the stronger is the evidence that H_0 is false and should be rejected. Unlike the P-value, the q-value takes into account other tests carried out at the same time. The idea is that, by choosing to reject H_0 only if the q-value is 0.05 or less, we reject the null hypothesis falsely in only 5% of tests. FDR and q-values are a more powerful approach to dealing with multiple comparisons, and we expect their use to increase in biological applications over the next decade. Consult Benjamini and Hochberg (1995) or Storey and Tibshirani (2003) for more details.