CHAPTER 13

Experiments and Observational Studies

Who gets good grades? And, more importantly, why? Is there something schools and parents could do to help weaker students improve their grades? Some people think they have an answer: music! No, not your iPod, but an instrument. In a study conducted at Mission Viejo High School, in California, researchers compared the scholastic performance of music students with that of non-music students. Guess what? The music students had a much higher overall grade point average than the non-music students, 3.59 to 2.91. Not only that: A whopping 16% of the music students had all A's compared with only 5% of the non-music students.

As a result of this study and others, many parent groups and educators pressed for expanded music programs in the nation's schools. They argued that the work ethic, discipline, and feeling of accomplishment fostered by learning to play an instrument also enhance a person's ability to succeed in school. They thought that involving more students in music would raise academic performance. What do you think? Does this study provide solid evidence? Or are there other possible explanations for the difference in grades? Is there any way to really prove such a conjecture?

Observational Studies

This research tried to show an association between music education and grades. But it wasn't a survey. Nor did it assign students to get music education. Instead, it simply observed students "in the wild," recording the choices they made and the outcome. Such studies are called **observational studies.** In observational studies, researchers don't *assign* choices; they simply observe them. In addition, this was a **retrospective study,** because researchers first identified subjects who studied music and then collected data on their past grades.

What's wrong with concluding that music education causes good grades? One high school during one academic year may not be representative of the

whole United States. That's true, but the real problem is that the claim that music study *caused* higher grades depends on there being *no other differences* between the groups that could account for the differences in grades, and studying music was not the *only* difference between the two groups of students.

We can think of lots of lurking variables that might cause the groups to perform differently. Students who study music may have better work habits to start with, and this makes them successful in both music and course work. Music students may have more parental support (someone had to pay for all those lessons), and that support may have enhanced their academic performance, too. Maybe they came from wealthier homes and had other advantages. Or it could be that smarter kids just like to play musical instruments.

For rare illnesses, it's not practical to draw a large enough sample to see many ill respondents, so the only option remaining is to develop retrospective data. For example, researchers can interview those who have become ill. The likely causes of both legionnaires' disease and HIV were initially identified from such retrospective studies of the small populations who were initially infected. But to confirm the causes, researchers needed laboratory-based experiments.

Observational studies are valuable for discovering trends and possible relationships. They are used widely in public health and marketing. Observational studies that try to discover variables related to rare outcomes, such as specific diseases, are often retrospective. They first identify people with the disease and then look into their history and heritage in search of things that may be related to their condition. But retrospective studies have a restricted view of the world because they are usually restricted to a small part of the entire population. And because retrospective records are based on historical data, they can have errors. (Do you recall *exactly* what you ate even yesterday? How about last Wednesday?)

A somewhat better approach is to observe individuals over time, recording the variables of interest and ultimately seeing how things turn out. For example, we might start by selecting young students who have not begun music lessons. We could then track their academic performance over several years, comparing those who later choose to study music with those who do not. Identifying subjects in advance and collecting data as events unfold would make this a **prospective study.**

Although an observational study may identify important variables related to the outcome we are interested in, there is no guarantee that we have found the right or the most important related variables. Students who choose to study an instrument might still differ from the others in some important way that we failed to observe. It may be this difference—whether we know what it is or not—rather than music itself that leads to better grades. It's just not possible for observational studies, whether prospective or retrospective, to demonstrate a causal relationship.

FOR EXAMPLE Designing an observational study

In early 2007, a larger-than-usual number of cats and dogs developed kidney failure; many died. Initially, researchers didn't know why, so they used an observational study to investigate.

Question: Suppose you were called on to plan a study seeking the cause of this problem. Would your design be retrospective or prospective? Explain why.

I would use a retrospective observational study. Even though the incidence of disease was higher than usual, it was still rare. Surveying all pets would have been impractical. Instead, it makes sense to locate some who were sick and ask about their diets, exposure to toxins, and other possible causes.

Randomized, Comparative Experiments

Experimental design was advanced in the 19th century by work in psychophysics by Gustav Fechner (1801–1887), the founder of experimental psychology. Fechner designed ingenious experiments that exhibited many of the features of modern designed experiments. Fechner was careful to control for the effects of factors that might affect his results. For example, in his 1860 book Elemente der Psychophysik *he cautioned readers to group experiment trials together to minimize the possible effects of time of day and fatigue.*

An Experiment:
Manipulates **the factor levels to create treatments.**
Randomly assigns **subjects to these treatment levels.**
Compares **the responses of the subject groups across treatment levels.**

"He that leaves nothing to chance will do few things ill, but he will do very few things."

—Lord Halifax (1633–1695)

Is it *ever* possible to get convincing evidence of a cause-and-effect relationship? Well, yes it is, but we would have to take a different approach. We could take a group of third graders, randomly assign half to take music lessons, and forbid the other half to do so. Then we could compare their grades several years later. This kind of study design is called an **experiment.**

An experiment requires a **random assignment** of subjects to treatments. Only an experiment can justify a claim like "Music lessons cause higher grades." Questions such as "Does taking vitamin C reduce the chance of getting a cold?" and "Does working with computers improve performance in Statistics class?" and "Is this drug a safe and effective treatment for that disease?" require a designed experiment to establish cause and effect.

Experiments study the relationship between two or more variables. An experimenter must identify at least one explanatory variable, called a **factor,** to manipulate and at least one **response variable** to measure. What distinguishes an experiment from other types of investigation is that the experimenter actively and deliberately manipulates the factors to control the details of the possible treatments, and assigns the subjects to those treatments *at random.* The experimenter then observes the response variable and *compares* responses for different groups of subjects who have been treated differently. For example, we might design an experiment to see whether the amount of sleep and exercise you get affects your performance.

The individuals on whom or which we experiment are known by a variety of terms. Humans who are experimented on are commonly called **subjects** or **participants.** Other individuals (rats, days, petri dishes of bacteria) are commonly referred to by the more generic term **experimental unit.** When we recruit subjects for our sleep deprivation experiment by advertising in Statistics class, we'll probably have better luck if we invite them to be participants than if we advertise that we need experimental units.

The specific values that the experimenter chooses for a factor are called the **levels** of the factor. We might assign our participants to sleep for 4, 6, or 8 hours. Often there are several factors at a variety of levels. (Our subjects will also be assigned to a treadmill for 0 or 30 minutes.) The combination of specific levels from all the factors that an experimental unit receives is known as its **treatment.** (Our subjects could have any one of six different treatments—three sleep levels, each at two exercise levels.)

How should we assign our participants to these treatments? Some students prefer 4 hours of sleep, while others need 8. Some exercise regularly; others are couch potatoes. Should we let the students choose the treatments they'd prefer? No. That would not be a good idea. To have any hope of drawing a fair conclusion, we must assign our participants to their treatments *at random*.

It may be obvious to you that we shouldn't let the students choose the treatment they'd prefer, but the need for random assignment is a lesson that was once hard for some to accept. For example, physicians might naturally prefer to assign patients to the therapy that they think best rather than have a random element such as a coin flip determine the treatment. But we've known for more than a century that for the results of an experiment to be valid, we must use deliberate randomization.

The Women's Health Initiative is a major 15-year research program funded by the National Institutes of Health to address the most common causes of death, disability, and poor quality of life in older women. It consists of both an observational study with more than 93,000 participants and several randomized comparative experiments. The goals of this study include

- giving reliable estimates of the extent to which known risk factors predict heart disease, cancers, and fractures;

No drug can be sold in the United States without first showing, in a suitably designed experiment approved by the Food and Drug Administration (FDA), that it's safe and effective. The small print on the booklet that comes with many prescription drugs usually describes the outcomes of that experiment.

- identifying "new" risk factors for these and other diseases in women;
- comparing risk factors, presence of disease at the start of the study, and new occurrences of disease during the study across all study components; and
- creating a future resource to identify biological indicators of disease, especially substances and factors found in blood.

That is, the study seeks to identify possible risk factors and assess how serious they might be. It seeks to build up data that might be checked retrospectively as the women in the study continue to be followed. There would be no way to find out these things with an experiment because the task includes identifying new risk factors. If we don't know those risk factors, we could never control them as factors in an experiment.

By contrast, one of the clinical trials (randomized experiments) that received much press attention randomly assigned postmenopausal women to take either hormone replacement therapy or an inactive pill. The results published in 2002 and 2004 concluded that hormone replacement with estrogen carried increased risks of stroke.

FOR EXAMPLE **Determining the treatments and response variable**

Recap: In 2007, deaths of a large number of pet dogs and cats were ultimately traced to contamination of some brands of pet food. The manufacturer now claims that the food is safe, but before it can be released, it must be tested.

Question: In an experiment to test whether the food is now safe for dogs to eat,[1] what would be the treatments and what would be the response variable?

The treatments would be ordinary-size portions of two dog foods: the new one from the company (the test food) and one that I was certain was safe (perhaps prepared in my kitchen or laboratory). The response would be a veterinarian's assessment of the health of the test animals.

The Four Principles of Experimental Design

A S **Video: An Industrial Experiment.** Manufacturers often use designed experiments to help them perfect new products. Watch this video about one such experiment.

1. **Control.** We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups. For human subjects, we try to treat them alike. However, there is always a question of degree and practicality. Controlling extraneous sources of variation reduces the variability of the responses, making it easier to detect differences among the treatment groups.

   Making generalizations from the experiment to other levels of the controlled factor can be risky. For example, suppose we test two laundry detergents and carefully control the water temperature at 180°F. This would reduce the variation in our results due to water temperature, but what could we say about the detergents' performance in cold water? Not much. It would be hard to justify extrapolating the results to other temperatures.

   Although we control both experimental factors and other sources of variation, we think of them very differently. We control a factor by assigning subjects to different factor levels because we want to see how the response will change at those different levels. We control other sources of variation to *prevent* them from changing and affecting the response variable.

[1] It may disturb you (as it does us) to think of deliberately putting dogs at risk in this experiment, but in fact that is what is done. The risk is borne by a small number of dogs so that the far larger population of dogs can be kept safe.