1 Design of Experiments

In many datasets we have more than one variable and we wish to describe and explain the relationships between them. Often, we would like to establish a cause-and-effect relationship.

1.1 Observational Studies

The American Music Conference is an organization that promotes music education at all levels. On their website [http://www.amc-music.com/research_briefs.htm](http://www.amc-music.com/research_briefs.htm) they promote music education as having all sorts of benefits. For example, they quote a study performed at the University of Sarasota in which “middle school and high school students who participated in instrumental music scored significantly higher than their non-band peers in standardized tests”. Does this mean that if the availability of and participation in instrumental programs in a school is increased, standardized test scores would generally increase? The American Music Conference is at least suggesting that this is true. They are attempting to “explain” the variation in test scores by the variation in music participation. The problem with that conclusion is that there might be other factors that cause the higher test scores of the band students. For example, students who play in bands are more likely to come from schools with more financial resources. They are also more likely to be in families that are actively involved in their education. It might be that music participation and higher test scores are a result of these variables. Such variables are often called lurking variables. A lurking variable is any variable that is not measured or accounted for but that has a significant effect on the relationship of the variables in the study.

The Sarasota study described above is an observational study. In such a study, the researcher simply observes the values of the relevant variables on the individuals studied. But as we saw above, an observational study can never definitively establish a causal relationship between two variables. This problem typically bedevils the analysis of data concerning health and medical treatment. The long process of establishing the relationship between smoking and lung cancer is a classic example. In 1957, the Joint Report of the Study Group on Smoking and Health concluded (in *Science*, vol. 125, pages 1129–1133) that smoking is an important health hazard because it causes an increased risk for lung cancer. However for many years after that the tobacco industry denied this claim. One of their principal arguments is that the data indicating this relationship came from observational studies. (Indeed, the data in the Joint Report came from 16 independent observational studies.) For example, the report documented that one out of every ten males who smoked at least two packs a day died of lung cancer. but only one out of every 275 males who did not smoke died of lung cancer. Data such as this falls short of establishing a cause-and-effect relationship however as there might be other variables that increase both one’s disposition to smoke and susceptibility to lung cancer.

Observational studies are useful for identifying possible relationships and also simply for describing relationships that exist. But they can never establish that there is a causal relationship between variables. Using observational studies in this way is analogous to using convenience samples to make
inferences about a population. There are some observational studies that are better than others however. The music study described above is a retrospective study. That is the researchers identified the subjects and then recorded information about past music behavior and grades. A prospective study is one in which the researcher identifies the subjects and then records variables over a period of time. A prospective study usually has a greater chance of identifying relevant possible “lurking” variables so as to rule them out as explanations for a possible relationship.

One of the most ambitious and scientifically important prospective observational studies has been the Framingham Heart Study. In 1948, researchers identified a sample of 5,209 adults in the town of Framingham, Massachusetts (a town about 25 miles west of Boston). The researchers tracked the lifestyle choices and medical records of these individuals for the rest of their lives. In fact the study continues to this day with the 1,110 individuals who are still living. The researchers have also added to the study 5,100 children of original study participants. There is no question that the Framingham Heart Study has led to a much greater understanding of what causes heart disease although it is “only” an observational study. For example, it is this study that gave researchers the first convincing data that smoking can cause high blood pressure. The website of the study http://www.nhlbi.nih.gov/about/framingham/ gives a wealth of information about the study and about cardiovascular health.

1.2 Randomized Comparative Experiments

If an observational study falls short of establishing a causal relationship and even an expensive well-designed prospective observational study cannot identify all possible lurking variables, can we ever prove such a relationship?

The “gold standard” for establishing a cause and effect relationship between two variables is the randomized comparative experiment. In an experiment, we want to study the relationship between two or more variables. At least one variable is an explanatory variable and the value of the variable can be controlled or manipulated. At least one variable is a response variable. The experimenter has access to a certain set of experimental units (subjects, individuals, cases), sets various values of the explanatory variables to create a treatment, and records the values of the response variables.

It is important first of all that an experiment be comparative. If we are attempting to establish that music participation increases grades, we cannot simply look at participators. We need to compare the achievement level of participators to those who do not participate. Many educational studies fall short of this standard. A school might introduce a new curriculum in mathematics and measure the test scores of the students at the end of the year. However the school cannot make the case that the test scores are a result of the new curriculum — the students might have achieved the same level with any curriculum.

In a randomized experiment we assign the individuals to the various treatments at random. For example, if we took 100 fifth graders and randomly chose 50 of them to be in the band and 50 of them not to receive any music instruction, we could begin to believe that differences in their test scores could be explained by the different treatments.

Example 1.2.1. Patients undergoing certain kinds of eye surgery are likely to experience serious post-operative pain. Researchers were interested in the question of whether giving acetaminophin to the patients before they experienced any pain would substantially reduce the subsequent pain
and the further need for analgesics. One group received acetaminophin before the surgery but no pain medicine after the surgery. A second group received no pain medicine before the surgery and acetaminophin after the surgery. And the third group received no acetaminophin either before or after the surgery. Sixty subjects were used and 20 subjects were assigned at random to each group. (Soltani, Hashemi, and Babaei, Journal of Research in Medical Sciences, March and April 2007; vol. 12, No 2.)

In Example 1.2.1, the goal of random assignment is to construct groups that are likely to be representative of the whole pool of subjects. If the assignment were left to the surgeons, for example, it might be the case that surgeons would give more pain medication to certain types of patients and therefore we wouldn’t be able to attribute the different results to the different treatments.

**Example 1.2.2.** The R dataset `chickwts` gives the weights of chicks who were fed six different diets over a period of time. The experimenter was attempting to determine which chicken feed caused the greatest weight gain. Feed is the explanatory variable and there were six treatments (six different feeds). Weight is the response variable. The first step in designing such an experiment is to assign baby chicks at random to the six different feed groups. If we allow the experimenter to choose which chicks receive which feed, she might unconsciously (or consciously) construct treatment groups that are unequal to start.

Student (W.S. Gosset) was one of the researchers in the early part of the twentieth century who realized the importance of randomization. One of his influential papers analyzed a large scale study that was to compare the nutritional effects of pasteurized and unpasteurized milk. In the Spring of 1930, 20,000 school children participated in the study. Of these, 5,000 received pasteurized milk each day, 5,000 received unpasteurized milk, and 10,000 did not receive milk at all. The weight and height of each student was recorded both before and after the trial. Student analyzed the way in which students were assigned to the three experimental treatments. There were 67 schools involved and in each school about half the students were in the control group and half received milk. However each school received only one kind of milk, pasteurized or unpasteurized. This was the first sort of bias that Student found — he was not convinced that the schools that received pasteurized milk were comparable to those that received unpasteurized milk. A more important difficulty was the way in which students were assigned either to the control or milk group within a school. The students were assigned at random initially, but teachers were given freedom to adjust the assignments if it seemed to them that the two groups were not comparable to each other in weight and height. In fact Student showed that this freedom on the part of teachers to assign subjects to groups resulted in a systematic difference between the groups in initial weight and height. The control groups were taller and heavier on average than those in the milk groups. Student conjectured that teachers unconsciously favored giving milk to the more undernourished students.

Of course assigning subjects to treatments at random does not ensure that the experimental groups are alike in all relevant ways. Just as we were subjected to sampling error when choosing a random sample from a population, we can have variation in the groups due to the chance mechanism alone. But assigning subjects at random will allow us to make probabilistic statements about the likelihood of such error just as we were able to make confidence intervals for parameters based on our analysis of sampling error that might arise in random sampling.
Randomized assignment and random samples

We assign subjects to treatments at random so that the various treatment groups will be similar with respect to the variables that we do not control. That is, we would like the experimental groups to be representative of the whole group of subjects. We choose a random sample from a population for a similar reason. We hope that the random sample is representative of a larger population. Ideally, we would like both kinds of randomness in our experiments. Not only do we ensure that the subjects are assigned at random to treatments, but we would like the subjects to be chosen at random from a larger population. If this is true, we could more easily justify generalizing our experimental results to a larger population than the immediate subject pool. However that is almost never the case. In the pain study of Example 1.2.1, the subjects were simply all those persons who were operated on at a given clinic in a given period of time. This issue is particularly important if we try to generalize the conclusions of an experiment to a larger population.

Example 1.2.3. The author of these notes participated in a study to investigate how people make probabilistic judgments in situations for which they do not have much data. (Default Probabilities, Osherson, Smith, Stob, and Wilkie, Cognitive Science, (15), 1991, 251–270.) Subjects were placed in various experimental groups at random. However the subjects were not chosen at random from any particular population. Indeed every subject was an undergraduate in an introductory psychology course at the University of Michigan or Massachusetts Institute of Technology. It is difficult to make an argument that the results of the paper would generalize to the population of all undergraduates in the United States let alone to the population of all adults. The MIT students in particular seemed to have a different set of strategies for dealing with probabilistic arguments.

Other features of a good experiment

In our analysis of simple random sampling from a population, we saw again and again the importance of large samples in getting precise estimates of our parameters. Analogously, if we are to measure precisely the effect of a treatment, we would like many individuals in each treatment group. This principle is known as replication. With a small number of individuals, it might be difficult to determine whether the differences in response are due only to the treatments or whether they reflect the natural variation in individuals. The chickwts data illustrate the issue. Figure 1.1 plots the weights of the six different treatment groups of chicks. While there is definitely some
variation between the groups, there is also considerable variation within each group. Chicks fed meatmeal, for example, have weights spanning most of the range of the entire experimental group. It is probably the case that the small difference between the linseed and soybean groups is due to the particular chicks in the groups rather than due to the feed. More chickens in each group would help us resolve this issue however.

In most good experiments one of the treatments is a control. A control generally means a treatment that is a baseline or status quo treatment. In an educational experiment, the control group might receive the standard curriculum while another group is receiving the supposed improved curriculum. In a medical experiment, the control group might receive the generally accepted treatment (or no treatment at all if ethical) while another group receives a new drug. In Example 1.2.1 the group that received no pre-pain medication is referred to as the control group. The goal of a control group is to establish a baseline to which to compare the new or changed treatment.

Often the control is a placebo. A placebo is a “treatment” that is really no treatment at all but looks like a treatment from the point of view of the subject. In Example 1.2.1 all subjects received pills both before and after surgery. But some of these pills contained no acetaminophen and were inert. Placebos are given to ensure that the placebo effect is measurable. The placebo effect is the tendency for experimental subjects to be affected by the treatment even if it has no content. The need for control groups and placebos is highlighted by the next famous example.

Example 1.2.4. During the period 1927-1932, researchers conducted a large-scale study of industrial efficiency at the Hawthorne Plant of the Western Electric Company in Cicero, IL. The researchers were interested in how physical and environmental features (e.g., lighting) affected worker productivity and satisfaction. Researchers found that no matter what the experimental conditions were, productivity tended to improve. Workers participating in the experiment tended to work harder and better to satisfy those persons who were experimenting on them. This feature of human experimentation — that the experimentation itself changes behavior whatever the treatment — is now called the Hawthorne Effect. (It is now generally accepted that the extent of the Hawthorne Effect in the original experiments have been significantly overstated by the gazillions of undergraduate psychology textbooks that refer to it. But the name remains and it makes a nice story as well as a plausible cautionary tale!)

Another feature which helps to ensure that the differences in treatments are due to the treatments themselves is blinding. An experiment is blind if the subjects do not know which treatment group that they are in. In Example 1.2.1 no subject knew whether they were receiving acetaminophen or a placebo. It is plausible that a subject knowing they receive a placebo would have a different (subjective) estimate of pain than one who thought that they might be receiving acetaminophen. An experiment is double-blind if the person administering the treatment also does not know which treatment is being administered. This prevents the researcher from treating the groups differently. It is not always possible or ethical to make an experiment blind or double-blind. But when possible, blinding helps to ensure that the differences between treatments are due to the treatments which is always the goal in experimentation.
1.3 Blocking

If the experimental subjects are identical, it does not matter which is assigned to which treatment. The differences in the response variable are likely to be the result of the differences in treatment. The subjects are not usually identical however or at least cannot be treated identically. So we would like to know that the differences in the response variable are due to the differences in the explanatory variable and not any systematic differences in subjects. Randomization is one tool that we use to distribute such differences equally across the treatments. In most cases however, our experimental units are not identical or our experiment itself introduces a systematic difference in the units that is due to something other than the treatment variable. This leads to the notion of blocking which we illustrate with a classic example.

R.A. Fisher was one of the key early figures in developing the principles of good experimental design. He did much of this while working at Rothamsted Experimental Station on agricultural experiments. He studied closely data from experiments that were attempting to establish such things as the effects of fertilizer on yield. Suppose that we have three unimaginatively named fertilizers $A$, $B$, $C$. We could divide the plot of land that we are using as in the first diagram of Figure 1.2. But it might be the case that the further north in the plot, the better the soil conditions. In that case, the variation in yield might be better explained (or at least partially explained) by the location of the plot rather than by fertilizer. In this example, we would say that the effects of northernness and fertilizer are confounded, meaning simply that we cannot separate them given the data of the experiment at hand. To separate out the effect of northernness from that of fertilizer, we could instead divide the patch using the second diagram in figure 1.2. Of course there still might be variations in the soil conditions across the three fertilizers. But we would at least be able to measure the effect of northernness separately from that of fertilizer. In this example, “northernness” is a blocking variable and our goal is to isolate the variability attributable to northernness so that we can see the differences between the fertilizers more clearly.

In a medical experiment it is often the case that gender or age are used as blocking variables. Obviously, we cannot assign individuals to the various levels of these variables at random but it is plausible that in certain circumstances gender or age can have a significant effect on the response. If so, it would be useful to design an experiment that allows us to separate out the effects of, say, gender and the treatment.
When using a blocking variable, it is important to continue to honor the principle of randomization. Suppose for example that we use gender as a blocking variable in a medical experiment comparing two treatments. The ideal experimental design would be to take a group of females and assign them at random to the two treatments and similarly for the group of males. That is, we should randomize the treatments within the blocks. The resulting experiment is usually called a **randomized block design**. It is not completely randomized because subjects in one block cannot be assigned to another but within a block it is randomized.

A special case of blocking is known as a **matched pair** design. In such an experiment, there are just two observations in each block (one for each of two treatments). In his 1908 paper, Student analyzed earlier published data from such an experiment. That data is in the R dataframe `sleep`. The two different treatments were two different soporifics (sleeping drugs). There was no control treatment. The response variable was the number of extra hours of sleep gained by the subject over his “normal” sleep. There were just 10 subjects and each subject took both drugs (on different nights). Thus each subject was a block and there was one observation on each treatment in each block. Student then compared the difference in the two drugs on each patient. Using the individuals as blocks served to help Student to decide what part of the variation in the response could be explained by the normal variation between individuals and what could be attributed to the drugs themselves.

In educational experiments, matched pairs are often constructed by finding two students who are very similar in baseline academic performance. Then it is hoped that the differences between these students at the end of the experiment are the result of the different treatments.

It is important to remember that block designs are not an alternative to randomization. Indeed, it is very important that we randomize the assignment to treatments within every block for the same reasons that randomization is important when we have no blocking variable. Identifying blocking variables is simply acknowledging that there are variables on which the treatments may systematically differ.

### 1.4 Experimental Design

In the above sections, we have introduced the three key features of a good experimental design — randomization, replication, blocking. We’ve illustrated these principles in the case that we have just one explanatory variable with just a few levels. These principles can be extended to situations with more than one explanatory variable however. In this book, we will not investigate the problem of inference for such situations or discuss in detail the issues of experimental design in these cases. In this section, we look at one example of extending these principles to experiments involving more than one explanatory variable.

**Example 1.4.1.** The R dataframe `ToothGrowth` contains the results of an experiment performed on Guinea Pigs to determine the effect of Vitamin C on tooth growth. There were two treatment variables, the dose of Vitamin C, and the delivery method of the Vitamin C. The dose variable had three levels (.5, 1, and 2 mg) and the delivery method was by either orange juice or ascorbic acid. There were 10 guinea pigs given each of the six treatments.

The plot below (using `coplot()`) shows the differences between the two delivery methods and the various doses levels.
It appears that both the delivery method and the dose have some effect on tooth growth.

Both the principles of randomization and replication extend to experiments with more than one explanatory variable. In Example 1.4.1 for example, it is apparent that the 60 guinea pigs should have been assigned at random to the six different treatments. And it also is clear that there should have been enough guinea pigs in each treatment so that the natural variation from pig to pig can be accounted for.

No blocking variables are described in the tooth growth study but it is often the case that natural blocking variables can be identified. For example, in the tooth growth study, it might not have been possible for the same technician to have recorded all the measurements. In that case, it would not be a good idea for one technician to make all the measurements for the orange juice treatment while another technician makes all the measurements for the ascorbic acid treatment. The blocking variable would be the technician and we would attempt to randomize assignment within treatment. Since there were 10 guinea pigs in each of the 6 treatments, two technicians could each measure 5 guinea pigs in each treatment.