Chapter 1
Statistics and Causality

What Is Statistics?

Question: What's the difference between accountants and statisticians?
Answer: Well, they both work with numbers, but statisticians just don't have the personality to be accountants.

Such is the stereotype of statisticians and statistics. Dull, plodding, and concerned with the tedious bean-counting enterprise of compiling numbers and tables and graphs on topics nobody much cares about. Nothing could be further from the truth. Well, okay, statisticians *are* dull; but statistics is one of the most exciting disciplines of all. Like astronomy, it's an investigation of the unknown—and, possibly, *unknowable*—world that's largely invisible to the naked eye. But this world is the one right under our noses: in terms of the subject of this book, it consists of human beings and their health. In this first chapter, we will consider what statistics is and why it is essential to the medical enterprise, and to science in general. Here, we define the science of statistics and relate it to real-world medical problems. Medical research is typically concerned with cause-and-effect relationships. The causes of disease or of health problems are important, as are the causal effects of treatments on medical outcomes. Therefore, we also discuss in this chapter the notion of a causal effect, and we ponder the conditions necessary for inferring causality in research.

What Statistics Is

Statistics is the science of converting data into evidence. *Data* constitute the raw material of statistics. They consist of numbers, letters, or special characters representing measurements of properties made on a collection of cases. Cases are the units of analysis in our study. Cases are usually people, but they could be days of

A. DeMaris and S.H. Selman, *Converting Data into Evidence: A Statistics Primer for the Medical Practitioner*, DOI 10.1007/978-1-4614-7792-1_1,

the week, organizations, nations or, in meta-analyses, other published studies. *Evidence* refers to information pertinent to judging the truth or falsehood of an assertion. The heart of statistics is called *inferential* statistics. It's concerned with making inferences about some population of cases. To do that, it uses a sample drawn from that population of cases and studies *it* rather than the entire population. On the basis of findings from the sample, we estimate some characteristic of a population or we judge the plausibility of statements made about the population. Let's take an example.

An Example

A frequent interest in medical research is HIV transmission and the course of the disease for those who are so infected (see, for example, Bendavid et al. 2012; Paton et al. 2012). Suppose a team of medical researchers is interested in the association between recreational intravenous drug use (IVDU) and contracting HIV in the USA. They believe that needle sharing is the prime means of transmission of this disease among the IVDU population. So they want to estimate the proportion of that population who is involved in needle sharing, for one thing. Then they want to test the hypothesis that needle sharing is a risk factor for becoming HIV positive (HIV+). But if they find that needle sharing is, in fact, associated with an elevated risk for HIV+, they want to ensure that it is the practice of sharing needles that is the "driver" of this association. That is, they need to rule out the possibility that it is some other risky behavior associated with sharing needles that is actually causing the association. Examples of other risky behaviors possibly associated with both IVDU and needle sharing are having unprotected sex, having sex with multiple partners, poor hygiene practices, and so forth.

This research problem presents several dilemmas. First, the population of interest is *all recreational IV drug users in the USA*. Now, what do you think the chances are of finding that population, let alone studying it? That's right—zip. Most users would not admit to having a drug habit, so we're unlikely to get very far surveying the USA population and asking people to self-identify as recreational IV drug users. So our let's say our team manages to recruit a sample of drug users, perhaps through a newspaper or magazine advertisement offering financial remuneration for taking part in a study. They find that 50 % of the sample of IV drug users share needles with other users. At this point the researchers would like to use this figure as their estimate of the proportion of all IV drug users in the USA who share needles. How should they proceed? Let's recognize, first, that the population proportion in question is a summary measure that statisticians refer to as a *parameter*. A parameter is just a summary statistic measuring some aspect of a population. Second, the parameter is unknown, and, in fact, *unknowable*. It's not possible to measure it directly, even though it exists "out there," somewhere. The best the team can do is to estimate it and then figure out how that estimate relates to the actual parameter value. We will spend much of this first part of the book on how this is accomplished.

Next, in order to test the primary hypothesis about needle sharing being a cause of HIV+ status, there has to be a comparison group of non-IV drug users. These individuals are much easier to find, since most people don't engage in IVDU. Let's say the team also recruits a control sample of such individuals, matched with the IVDU group on gender, age, race, and education. They then need to measure all the relevant variables. This includes the "mechanisms," aside from needle sharing, that they believe might be responsible for the IVDU-HIV+ association, i.e., having unprotected sex, having sex with multiple partners, quality of personal hygiene, and so forth. In order to fully evaluate the hypothesis, they will conduct a *multivariable analysis* (or *multivariate analysis*—these terms are used interchangeably). HIV+ status will be the primary *study endpoint* (or *response variable*), and needle sharing and the other risky behaviors will be the *explanatory variables* (or *regressors, predictors*, or *covariates*). The multivariable analysis will allow them to examine whether needle sharing is responsible for the (presumed) higher HIV+ rate among the IVDU vs. the non-IVDU group. It will also let them assess whether it is needle sharing, per se, rather than one of the other risky behaviors that is the driving factor. We will discuss multivariable statistical techniques in a later section of the book.

However, there are other complications to be dealt with. Suppose that some of the subjects of the study fail to provide answers to some of the questions? This creates the problem of *missing data*. We can simply discard these subjects from the study, but then we (a) lose all of the other information that they did provide and (b) introduce selection bias into the study because those who don't answer items are usually not just a random subset of the subjects. This means that those left in the sample are a select group—perhaps a more compliant type of individuals—and the results then will only apply to people of that type. One solution is that the researchers can *impute* the missing data and then still include the cases. Imputation is the practice of filling in the missing data with a value representing our best guess about what the missing value would be were it measured. The state of the art in imputation techniques is a procedure called *multiple imputation*. Multiple imputation will be covered later in the book in the chapter on advanced techniques.

The last major issue is that it's always possible that some characteristic that the researchers have not measured might be producing the association between needle sharing and HIV+ status. That is, it's not really needle sharing that elevates HIV+ risk. It's some unmeasured characteristic of individuals that also happens to be associated with needle sharing. An unmeasured characteristic that may be influencing one's results is often referred to as *unmeasured heterogeneity*. The term refers to the fact that the characteristic exhibits heterogeneity—i.e., variance—across individuals that is related to the variation in the study endpoint. The fact that it is unmeasured means that there is no easy way to control for it in our analyses. We will discuss this problem in greater detail later in this chapter. And we will see one possible statistical solution to this problem, called *fixed-effects regression modeling*, when we get to the advanced techniques chapter. In sum, statistics allows us to address research problems of the foregoing nature and provide answers to these kinds of complex questions that are posed routinely in research.

Populations and Samples

The population in any study is the *total collection of cases we want to make assertions about*. A "case" is the smallest element constituting a single "replication" of a treatment. Suppose, for example, that you are interested in the effect of diet on prostate-specific antigen (PSA). You suspect that a diet heavy in red meat contains carcinogens that raise the risk for prostate cancer. So you anticipate that a red-meat-rich diet will be associated with higher PSA levels. Suppose you have a sample of six men from each of two groups: a control group eating a balanced diet and a treatment group eating a diet overloaded with red meat. In this case, individual men are the cases, since each man eating a particular diet represents a replication of the "treatment." By "treatment," in this case, we mean diet, of which there are two treatment levels: balanced and red-meat rich. Who is the population here? The population we'd ideally like to be talking about is the entire population of adult males in the USA. So our 12 men constitute a sample from it.

Probability vs. Nonprobability Samples

Statisticians distinguish two major classes of samples: probability and nonprobability. A *probability sample* is one for which one can specify the probability that any member of the population will be selected into it. *Nonprobability samples* do not have this property. The best-known probability sample is a simple random sample or SRS. An SRS is one in which every member of the population has the same chance of being selected into the sample. For example, if the population consists of 50,000 units and we're drawing an SRS of 50 units from it, each population member has a $50/50{,}000 = 0.001$ chance of being selected. Probability samples provide results that can be generalized to the population. Nonprobability samples don't. In our diet study example, if the 12 men were randomly sampled from the population of interest, the results could be generalized to that population. Most likely, though, the 12 men were recruited via advertisement or by virtue of being part of a patient population. If the 12 men weren't sampled randomly "from" a known population, then what kind of population might they represent?

Sampling "to" a Population

Many samples in science are of the nonprobability type. What can we say about the "population" of interest, then? Some statisticians will tell you: nothing. But that implies that your sample is so unique, there's no one else who behaves or responds the same way to a treatment. That's not very realistic. Rather, what we can do with nonprobability sample results is use the characteristics of sample participants to

suggest a hypothetical population the results might be generalizable to. Much of the time in studies of this nature, the sample consists of volunteers responding to a newspaper ad announcing a clinical trial. In research involving the human body, one could, of course, argue that people are sufficiently similar biologically that the 12 men in the example above are representative of men in general. But statistically, at least, generalizing to a population requires sampling randomly from it. Another way to define the population, however, is to reason in the opposite direction. That is, whatever the manner in which the 12 men were recruited for this study, suppose we repeat that recruitment strategy and collect 12 men a second time. And suppose we repeat it, again, and collect a third group of 12 men. And then suppose we go on and on like this, collecting sample after sample of 12 men by repeating the recruitment strategy over and over, ad infinitum. Eventually, the entire collection of men accumulating from all of these samples could be considered the "population." And our original sample of 12 men can then be thought of as a random sample from *this* population. This has been termed "sampling *to* a population," as opposed to sampling *from* a population (DeMaris 2004), and is one way of defining a conceptual population that one's inferences might apply to.

Statistics and Causal Inference

The scientific enterprise is typically concerned with cause and effect. What causes elevated PSA levels, for example? Or, what causes prostate cancer? Or, what causes prostate cancer to develop sooner rather than later? Statistics can aid in making causal inferences. To understand its utility in this arena, however, we first have to define what we mean by "cause," or, more properly, a "causal effect." The reigning definition in contemporary science is due to two statisticians, Jerzy Neyman and Donald Rubin (West and Thoemmes 2010). The Neyman–Rubin causal paradigm is simple, mathematically elegant, and intuitive. We normally think of a cause as something that changes life's "trajectory" from what would have transpired were the cause not operating. The Neyman–Rubin paradigm simply puts this in mathematical terms.

A Mathematical Definition of "Causal Effect"

Employing, again, the diet-PSA example, suppose a man follows a balanced diet for some period of time. His PSA level measured after that period would be denoted Yc. And then suppose he were instead to follow a meat-heavy diet for the same period. Denote his PSA level after that as Yt. Notice that this scenario is *contrary to fact*. He can't follow both diets over the same period; he's either on one or the other. But suspend disbelief for a moment and suppose that's what he does. The causal effect of the steak diet on PSA is defined as: $Yt - Yc$. It is the boost in PSA

attributable to the steak diet. So if his PSA is 2.6 on the balanced diet vs. 4.3 on the steak diet, the causal effect of diet is 4.3 − 2.6 = 1.7, or the steak diet results in a boost in PSA level by 1.7.

If we were to apply this regimen to every man in the population and then average all of the (Yt − Yc) differences, we would have the *Average Causal Effect*, or ACE, of the steak diet on PSA. The ACE is often the *parameter* of interest in research. If the outcome of interest is a qualitative one, then the true causal effect is defined with a slightly different measure. So if the man in question has a 30 % chance of developing prostate cancer on the balanced diet, but a 60 % chance on the steak diet, the causal effect of a steak diet on the risk of cancer is 0.60/0.30 = 2. Or, a steak diet doubles the risk of cancer for this particular man. The number 2 is called the *relative risk* for cancer due to a steak, vs. a balanced, diet.

How Do We Estimate the ACE?

Because the ACE is contrary-to-fact, and therefore not measurable, how can we estimate it? It turns out that the ACE can be estimated in an unbiased fashion as the mean difference in PSA levels between men on a balanced vs. a meat diet in a study if a particular condition is met. The condition is referred to as the *ignorability* condition: the treatment assignment mechanism is ignorable if the potential outcomes (e.g., PSA levels) are independent of the treatment assignment "mechanism." What this means in practice, using our example, is that there is no a priori tendency for those in the steak-diet condition to have higher or lower PSA levels than men in the other condition *before the treatments are even applied*. The only way to ensure this is to *randomly assign* the men to the two diet conditions, and this is the hallmark of the clinical trial, or, for that matter, any experimental study. Random assignment to treatment groups ensures that, *on average*, treatment and control groups are exactly the same on all characteristics at the beginning of a study. In this manner, we are assured that the treatment effect is a true causal effect and is not an artifact of a latent *self-selection factor*. It is random assignment to treatment levels that provides researchers with the best vehicle for inferring causality.

Example of Latent Self-Selection

As an example of latent self-selection confounding causal inference in a study, regard Fig. 1.1, below. It shows one possible scenario that could occur in the absence of random assignment, such as if we simply study groups of men who have chosen each type of diet themselves.

The negative numbers represent inverse relationships. The "–0.75" on the curved arrow connecting health awareness with meat diet is a *correlation coefficient*. It means those with greater health awareness are less likely to be on a meat diet. They

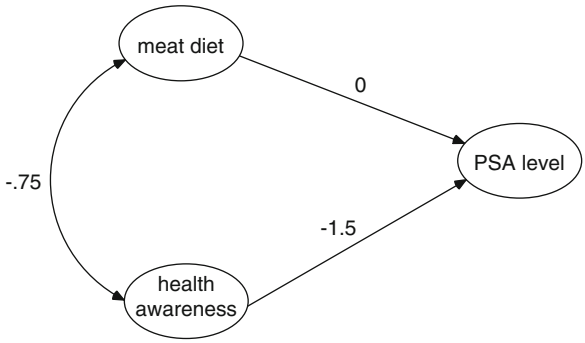

Fig. 1.1 Causal diagram of variables affecting PSA level

are probably men who lead healthy lifestyles that include moderate alcohol intake, nonsmoking, plenty of exercise, regular medical checkups, etc. The "−1.5" from health awareness to PSA levels is a *causal effect*. It means that health awareness leads to lower PSA levels. Simply looking at the difference in average PSA between the two groups of men while ignoring health awareness confounds the true relationship of diet to PSA. There might be no association of diet with PSA (shown by the "0" on that path in the diagram). But if health awareness is not "controlled" in the study, then the indirect link from meat diet to PSA level through health awareness will manifest itself as a positive "effect" of a meat diet on PSA level. This happens because ignoring health awareness is equivalent to multiplying together the two negative numbers: $(-0.75)\times(-1.5)=1.125$, and then adding the result to the path from meat diet to PSA level. This makes it appear that meat diet has a positive effect on PSA level: the "1.125" would appear to be the average PSA level difference between the men in the two groups. The take-home message here is simple: only random assignment to treatment conditions lets us confidently rule out latent selection factors as accounting for treatment effects in a study. In epidemiological and other observational—as opposed to experimental—studies, latent selection factors are an ever-present threat. They are typically countered by measuring any such selection factors ahead of time, and then statistically controlling for them when estimating causal effects. Under the right conditions, we can even eliminate *unmeasured* factors, as we shall see in the advanced techniques chapter. And we shall have more to say about statistical control, in general, later in this primer.

Internal vs. External Validity: A Conundrum

At this point, we have discussed the nature of causal effects, the advantages of random assignment to treatment conditions, and latent selection factors in nonexperimental studies. It is worth noting, as a final issue, that both experimental and nonexperimental studies have particular advantages and drawbacks. And both are regularly used in medical research. Statisticians speak of a study having

internal vs. external validity. *Internal validity* obtains to the extent that the treatment-group differences observed on a study endpoint strictly represent the causal effect of the treatment on the response variable (Singleton and Straits 2010). *External validity* obtains to the extent that the study's results can be generalized to a larger, known population. As we have noted, experimental studies, in which cases are randomly assigned to treatment groups, are ideal for estimating causal effects. The gold standard in this genre is the double-blind, placebo-controlled, clinical trial. Studies of this nature have a clear advantage in internal validity over nonexperimental studies. However, experimental studies may be deficient in external validity. For one thing, it may not be clear what population the study results are generalizable to. It is very rare—in fact, unheard of—for researchers to take a random sample of a patient population and then randomly assign sample members to treatment conditions. Patients are usually a "captive audience"; they are at hand by virtue of seeking treatment from a given clinic or hospital. Or they are recruited through advertisements for a clinical trial. As they don't typically represent a probability sample from a known population, it is not immediately clear what larger population they might represent. We can invoke the aforementioned notion of "sampling to a population" to justify a kind of generalizability. But the larger population the results might apply to is only hypothetical. A second factor that detracts from external validity is that, in actual clinical practice, patients are not randomly assigned to treatments. They elect to undergo certain treatments in consultation with their physician. Therefore, there is always an element of self-selection operating in the determination of which patients end up getting which treatments. This may lead to a different treatment outcome than if patients were randomly assigned to their treatments (Marcus et al. 2012). Thus, the pure causal effect observed in a clinical trial may not correspond perfectly to the real-world patient setting.

Nonexperimental studies often have an advantage in external validity. Many nonexperimental studies are based on probability sampling from a known population. Moreover, many follow patients after they have undergone treatments of their own choosing—on physician advice, of course. The disadvantage, as noted previously, is that nonexperimental study results can always be confounded by unmeasured heterogeneity. It is never possible to control for all possible patient characteristics that might affect the study results. Hence, nonexperimental studies often suffer from questions regarding their internal validity. We shall have much more to say about nonexperimental data analysis in subsequent chapters. In the meantime, the next chapter introduces techniques for summarizing the main features of a set of data. Understanding what your data "look like" is a first step in the research process.