About This Errata

This file contains changes made to the text of *The Book of Why: The New Science of Cause and Effect* by Judea Pearl and Dana Mackenzie.

Changes are pointed to with arrows in the margins which will make it easy for you to mark your own personal copy.

If you should discover additional corrections or needed clarification, please let us know (kaoru@cs.ucla.edu).

Many thanks.

→ Noted Errata through 5.20.2021
→ New errata, updated 4.24.2024

data—in other words, the cause-effect forces that operate in the environment and shape the data generated.

Side by side with this diagrammatic "language of knowledge," we also have a symbolic "language of queries" to express the questions we want answers to. For example, if we are interested in the effect of a drug (D) on lifespan (L), then our query might be written symbolically as: $P(L \mid do(D))$. The vertical line means "given that," so we are asking: what is the probability (P) that a typical patient would survive L years, given that he or she is made to take the drug ($do(D)$)? This question describes what epidemiologists would call an *intervention* or a *treatment* and corresponds to what we measure in a clinical trial. In many cases we may also wish to compare $P(L \mid do(D))$ with $P(L \mid do(\text{not-}D))$; the latter describes patients denied treatment, also called the "control" patients. The *do*-operator signifies that we are dealing with an intervention rather than a passive observation; classical statistics has nothing remotely similar to it.

We must invoke an intervention operator $do(D)$ to ensure that the observed change in Lifespan L is due to the drug itself and is not confounded with other factors that tend to shorten or lengthen life. If, instead of intervening, we let the patient himself decide whether to take the drug, those other factors might influence his decision, and lifespan differences between taking and not taking the drug would no longer be solely due to the drug. For example, suppose only those who were terminally ill took the drug. Such persons would surely differ from those who did not take the drug, and a comparison of the two groups would reflect differences in the severity of their disease rather than the effect of the drug. By contrast, forcing patients to take or refrain from taking the drug, regardless of preconditions, would wash away preexisting differences and provide a valid comparison.

Mathematically, we write the observed frequency of Lifespan L among patients who voluntarily take the drug as $P(L \mid D)$, which is the standard conditional probability used in statistical textbooks. This expression stands for the probability (P) of Lifespan L conditional on seeing the patient take Drug D. Note that $P(L \mid D)$ may be totally different from $P(L \mid do(D))$. This difference between seeing

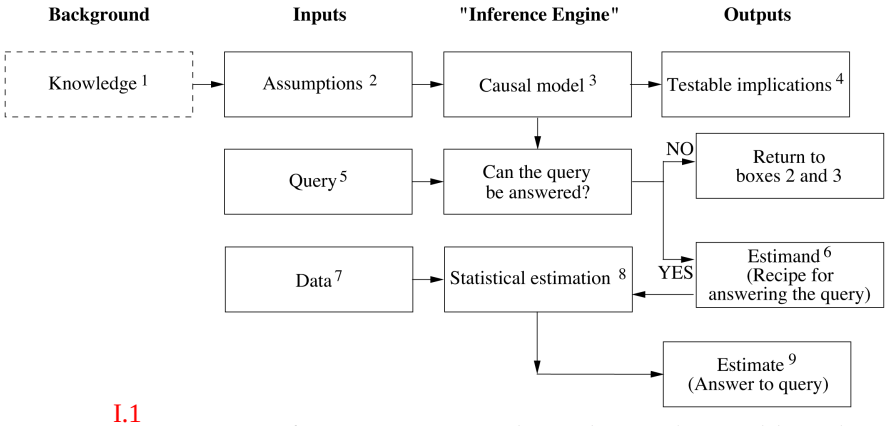

FIGURE I.1 How an "inference engine" combines data with causal knowledge to produce answers to queries of interest. The dashed box is not part of the engine but is required for building it. Arrows could also be drawn from boxes 4 and 9 to box 1, but I have opted to keep the diagram simple.

the Data input, it will use the recipe to produce an actual Estimate for the answer, along with statistical estimates of the amount of uncertainty in that estimate. This uncertainty reflects the limited size of the data set as well as possible measurement errors or missing data.

To dig more deeply into the chart, I have labeled the boxes 1 through 9, which I will annotate in the context of the query "What is the effect of Drug *D* on Lifespan *L*?"

1. "Knowledge" stands for traces of experience the reasoning agent has had in the past, including past observations, past actions, education, and cultural mores, that are deemed relevant to the query of interest. The dotted box around "Knowledge" indicates that it remains implicit in the mind of the agent and is not explicated formally in the model.
2. Scientific research always requires simplifying assumptions, that is, statements which the researcher deems worthy of making explicit on the basis of the available Knowledge. While most of the researcher's knowledge remains implicit in his or her brain, only Assumptions see the light of day and

5. Queries submitted to the inference engine are the scientific questions that we want to answer. They must be formulated in causal vocabulary. For example, what is $P(L \mid do(D))$? One of the main accomplishments of the Causal Revolution has been to make this language scientifically transparent as well as mathematically rigorous.
6. "Estimand" comes from Latin, meaning "that which is to be estimated." This is a statistical quantity to be estimated from the data that, once estimated, can legitimately represent the answer to our query. While written as a probability formula—for example, $P(L \mid D, Z) \times P(Z)$—it is in fact a recipe for answering the causal query from the type of data we have, once it has been certified by the engine.

   It's very important to realize that, contrary to traditional estimation in statistics, some queries may not be answerable under the current causal model, even after the collection of any amount of data. For example, if our model shows that both D and L depend on a third variable Z (say, the stage of a disease), and if we do not have any way to measure Z, then the query $P(L \mid do(D))$ cannot be answered. In that case it is a waste of time to collect data. Instead we need to go back and refine the model, either by adding new scientific knowledge that might allow us to estimate Z or by making simplifying assumptions (at the risk of being wrong)—for example, that the effect of Z on D is negligible.
7. Data are the ingredients that go into the estimand recipe. It is critical to realize that data are profoundly dumb about causal relationships. They tell us about quantities like $P(L \mid D)$ or $P(L \mid D, Z)$. It is the job of the estimand to tell us how to bake these statistical quantities into one expression that, based on the model assumptions, is logically equivalent to the causal query—say, $P(L \mid do(D))$.

   Notice that the whole notion of estimands and in fact the whole top part of Figure 1.1 does not exist in traditional methods of statistical analysis. There, the estimand and the query coincide. For example, if we are interested in the proportion

of people among those with Lifespan L who took the Drug D, we simply write this query as $P(D \mid L)$. The same quantity would be our estimand. This already specifies what proportions in the data need to be estimated and requires no causal knowledge. For this reason, some statisticians to this day find it extremely hard to understand why some knowledge lies outside the province of statistics and why data alone cannot make up for lack of scientific knowledge.

8. The estimate is what comes out of the oven. However, it is only approximate because of one other real-world fact about data: they are always only a finite sample from a theoretically infinite population. In our running example, the sample consists of the patients we choose to study. Even if we choose them at random, there is always some chance that the proportions measured in the sample are not representative of the proportions in the population at large. Fortunately, the discipline of statistics, nowadays empowered by advanced techniques of machine learning, gives us many, many ways to manage this uncertainty—maximum likelihood ~~estimators~~ parametric and semi-parametric models, ~~propensity scores, confidence intervals, significance tests, and so forth.~~ methods, and are often used to smooth the sparse data.
9. In the end, if our model is correct and our data are sufficient, we get an answer to our causal query, such as "Drug D increases the Lifespan L of diabetic Patients Z by 30 percent, plus or minus 20 percent." Hooray! The answer will also add to our scientific knowledge (box 1) and, if things did not go the way we expected, might suggest some improvements to our causal model (box 3).

This flowchart may look complicated at first, and you might wonder whether it is really necessary. Indeed, in our ordinary lives, we are somehow able to make causal judgments without consciously going through such a complicated process and certainly without resorting to the mathematics of probabilities and proportions. Our causal intuition alone is usually sufficient for handling the kind of uncertainty we find in household routines or even in our

They can call upon a tradition of thought about causation that goes back at least to Aristotle, and they can talk about causation without blushing or hiding it behind the label of "association."

However, in their effort to mathematize the concept of causation—itself a laudable idea—philosophers were too quick to commit to the only uncertainty-handling language they knew, the language of probability. They have for the most part gotten over this blunder in the past decade or so, but unfortunately similar ideas are being pursued in econometrics even now, under names like "Granger causality" and "vector ~~autocorrelation~~ autoregression."

Now I have a confession to make: I made the same mistake. I did not always put causality first and probability second. Quite the opposite! When I started working in artificial intelligence, in the early 1980s, I thought that uncertainty was the most important thing missing from AI. Moreover, I insisted that uncertainty be represented by probabilities. Thus, as I explain in Chapter 3, I developed an approach to reasoning under uncertainty, called Bayesian networks, that mimics how an idealized, decentralized brain might incorporate probabilities into its decisions. Given that we see certain facts, Bayesian networks can swiftly compute the likelihood that certain other facts are true or false. Not surprisingly, Bayesian networks caught on immediately in the AI community and even today are considered a leading paradigm in artificial intelligence for reasoning under uncertainty.

Though I am delighted with the ongoing success of Bayesian networks, they failed to bridge the gap between artificial and human intelligence. I'm sure you can figure out the missing ingredient: causality. True, causal ghosts were all over the place. The arrows invariably pointed from causes to effects, and practitioners often noted that diagnostic systems became unmanageable when the direction of the arrows was reversed. But for the most part we thought that this was a cultural habit, or an artifact of old thought patterns, not a central aspect of intelligent behavior.

At the time, I was so intoxicated with the power of probabilities that I considered causality a subservient concept, merely a convenience or a mental shorthand for expressing probabilistic dependencies and distinguishing relevant variables from irrelevant ones.

Another beautiful example of this can be found in his "Correlation and Causation" paper, from 1921, which asks how much a guinea pig's birth weight will be affected if it spends one more day in the womb. I would like to examine Wright's answer in some detail to enjoy the beauty of his method and to satisfy readers who would like to see how the mathematics of path analysis works.

Notice that we cannot answer Wright's question directly, because we can't weigh a guinea pig in the womb. What we can do, though, is compare the birth weights of guinea pigs that spend (say) sixty-six days gestating with those that spend sixty-seven days. Wright noted that the guinea pigs that spent a day longer in the womb weighed an average of 5.66 grams more at birth. So, one might naively suppose that a guinea pig embryo grows at 5.66 grams per day just before it is born.

"Wrong!" says Wright. The pups born later are usually born later for a reason: they have fewer litter mates. This means that they have had a more favorable environment for growth throughout the pregnancy. A pup with only two siblings, for instance, will already weigh more on day sixty-six than a pup with four siblings. Thus the difference in birth weights has two causes, and we want to disentangle them. How much of the 5.66 grams is due to spending an additional day in utero and how much is due to having fewer siblings to compete with?

Wright answered this question by setting up a path diagram (Figure 2.8). *X* represents the pup's birth weight. *Q* and *P* represent the two known causes of the birth weight: the length of gestation (*P*) and rate of growth in utero (*Q*). *L* represents litter size, which

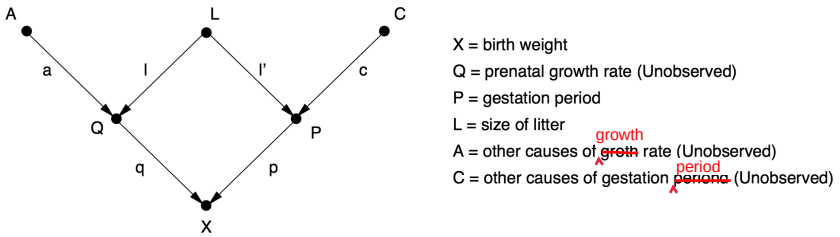

FIGURE 2.8. Causal (path) diagram for birth-weight example.

explain why it is harder; he took that as self-evident, proved that it is doable, and showed us how.

To appreciate the nature of the problem, let's look at ~~the~~ a slightly simplified version of an example he suggested ~~himself~~ in his posthumous paper of 1763. Imagine that we shoot a billiard ball on a table, making sure that it bounces many times so that we have no idea where it will end up. What is the probability that it will stop within *x* feet of the left-hand end of the table? If we know the length of the table and it is perfectly smooth and flat, this is a very easy question (Figure 3.2, top). For example, on a twelve-foot snooker table, the probability of the ball stopping within a foot of the end would be 1/12. On an eight-foot billiard table, the probability would be 1/8.

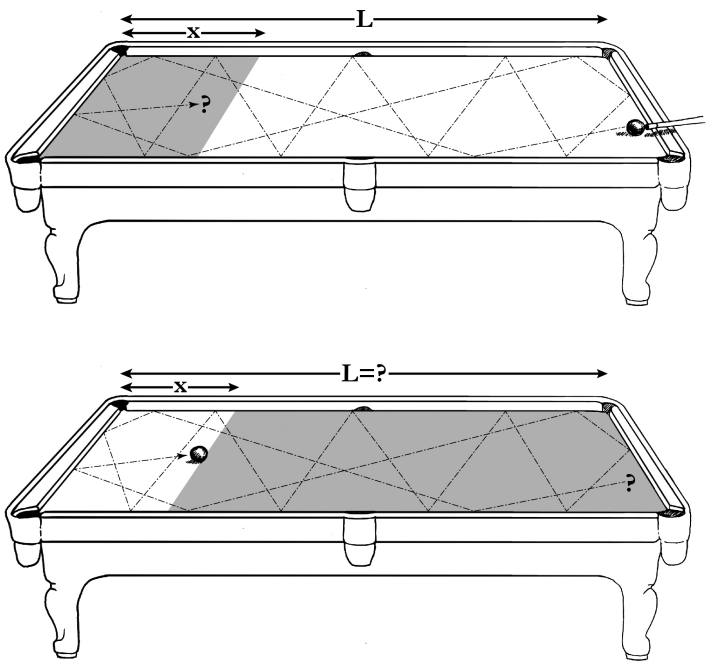

FIGURE 3.2. Thomas Bayes's pool table example. In the first version, a forward-probability question, we know the length of the table and want to calculate the probability of the ball stopping within *x* feet of the end. In the second, an inverse-probability question, we observe that the ball stopped *x* feet from the end and want to estimate the likelihood that the table's length is *L*. (*Source:* Drawing by Maayan Harel.)

not matter which came first, ordering tea or ordering scones. It only mattered which conditional probability we felt more capable of assessing. But the causal setting clarifies why we feel less comfortable assessing the "inverse probability," and Bayes's essay makes clear that this is exactly the sort of problem that interested him.

Suppose a forty-year-old woman gets a mammogram to check for breast cancer, and it comes back positive. The hypothesis, D (for "disease"), is that she has cancer. The evidence, T (for "test"), is the result of the mammogram. How strongly should she believe the hypothesis? Should she have surgery?

We can answer these questions by rewriting Bayes's rule ~~as follows:~~ in a more convenient form, using the concept of *odds* rather than probability:

$$\text{(Updated odds of } D\text{)} = \text{(likelihood ratio)} \times \text{(prior odds of } D\text{)}. \qquad (3.2)$$

~~(Updated probability of D) = $P(D \mid T)$ =~~

~~(likelihood ratio) × (prior probability of D)~~ ~~(3.2)~~

~~where the~~ ~~$P(T)$. It~~ ~~measures l~~ ~~ble with~~ ~~the disease~~ ~~herefore~~ ~~tells us tha~~ ~~f D by a~~ ~~fixed ratio~~

~~Let's de~~ ~~works.~~ ~~For a typic~~ ~~g breast~~ ~~cancer in t~~ ~~e'll use~~ ~~that as our~~

~~To con~~ ~~$(T \mid D)$~~ ~~and $P(T)$.~~ ~~y of the~~ ~~mammogr~~ ~~sitive if~~ ~~you have c~~ ~~ce Con-~~ ~~sortium (B~~ ~~ear-old~~ ~~women is~~

~~The de~~ ~~T, can~~ ~~come both~~ ~~patients~~ ~~who don't~~ ~~$P(T \mid D)$~~ ~~(the proba~~ ~~the dis-~~ ~~ease) and $P(T \mid \sim D)$ (the probability of a positive test among those~~

The prior odds of D are simply the ratio of the probability of having the disease to the probability of not having the disease, or $P(D)$ divided by $P(\sim D)$. The updated odds reflect the same ratio after the test is performed and the outcome is known: $P(D \mid T)$ divided by $P(\sim D \mid T)$. The new term in equation 3.2 is the "likelihood ratio," which is given by $P(T \mid D)$ divided by $P(T \mid \sim D)$. In words, the likelihood ratio equals the true positive rate of the test divided by the false positive rate. Equation 3.2 tells us that the new evidence augments the odds of D by a constant, which does not depend on the prevalence of the disease.

Let's do an example to see how this convenient formula works in the case of a breast cancer examination. To compute the likelihood ratio, we need to divide $P(T \mid D)$ by $P(T \mid \sim D)$, two readily available quantities. In the medical context, $P(T \mid D)$ is the sensitivity of the mammogram—the probability that it will come back positive if you have cancer. According to the Breast Cancer Surveillance Consurtium (BCSC), the sensitivity of mammograms for forty-year-old women is 73 percent.

The denominator, $P(T \mid \sim D)$, gives the percentage

of patients who receive a positive test result, *T*, even though they do not have the disease, ~D. This percentage is known as the false positive rate. According to the BCSC, the false positive rate for forty-year-old women is about 12 percent. This means, then, that the likelihood ratio is 73 percent divided by 12 percent, or slightly more than 6. (We will round off to 6 for simplicity.)

According to formula 3.2, we need one other piece of information to properly interpret the positive test result: the prior odds of having breast cancer. According to the BCSC, roughly 1 in 700 women at age 40 has breast cancer, so the odds of having cancer are (1/700) divided by (699/700) or 1:699. (Another way of saying this is that the odds against cancer are 699:1.)

Putting all of this information together, Bayes's rule tells us that the updated odds that the patient has breast cancer, after receiving a positive test result, are 6 × (1/699) ≈ 1:116. Note that the odds have increased by a factor of 6 (the likelihood ratio) but they are nevertheless exceedingly small. She still has less than a 1 percent chance of having cancer.

The conclusion is startling. I think that most forty-year-old women who have a positive mammogram would be astounded to learn that they still have less than a 1 percent chance of having breast cancer. Figure 3.3 might make the reason easier to understand. The tiny number of true positives (i.e., women with breast cancer who tested positive) is overwhelmed by the number of false positives. Our sense of surprise at this result comes from the common cognitive confusion between the forward probability and the inverse probability. The forward probability, which is well studied and thoroughly documented, is the probability of a positive test if you have cancer: 73 percent. But the probability that is needed for individual decision making is the inverse probability: the probability that you have cancer if the test is positive. In our example, this probability is less than 1 percent.

The conflict between our perception and reality partially explains the outcry when the US Preventive Services Task Force, in 2009, recommended that forty-year-old women should not get annual mammograms. The task force understood what many women

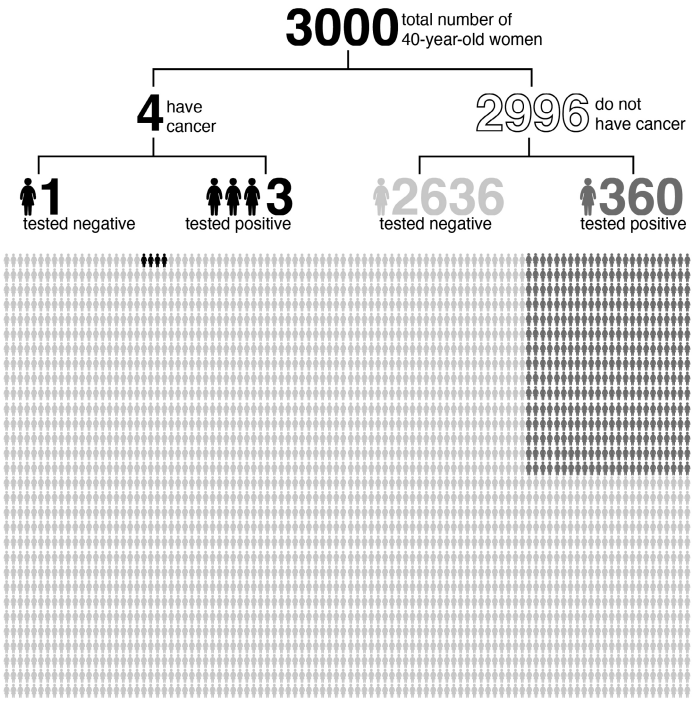

FIGURE 3.3. In this example, based on false-positive and false-negative rates provided by the Breast Cancer Surveillance Consortium, only 3 out of 363 forty-year-old women who test positive for breast cancer actually have the disease. (Proportions do not exactly match the text because of rounding.) (*Source:* Infographic by Maayan Harel.)

did not: a positive test at that age is way more likely to be a false alarm than to detect cancer, and many women were unnecessarily terrified (and getting unnecessary treatment) as a result.

However, the story would be very different if our patient had a gene that put her at high risk for breast cancer—say, a one-in-twenty chance ~~within the next year~~. Then a positive test would increase her odds from 1:19 to 6:19 (or increase the probability to 24 percent). ~~the probability to almost one in three.~~ For a woman in this situation, the chances that the test provides lifesaving information are much higher. That is why the task force continued to recommend annual mammograms for high-risk women.

This example shows that $P(disease \mid test)$ is not the same for everyone; it is context dependent. If you know that you are at high

merely signifies that we know the "forward" probability, *P*(*scones* | *tea*) or *P*(*test* | *disease*). Bayes's rule tells us how to reverse the procedure, ~~specifically~~ for example by multiplying the prior ~~probability~~ odds by a likelihood ratio.

Belief propagation formally works in exactly the same way whether the arrows are noncausal or causal. Nevertheless, you may have the intuitive feeling that we have done something more meaningful in the latter case than in the former. That is because our brains are endowed with special machinery for comprehending cause-effect relationships (such as cancer and mammograms). Not so for mere associations (such as tea and scones).

The next step after a two-node network with one link is, of course, a three-node network with two links, which I will call a "junction." These are the building blocks of all Bayesian networks (and causal networks as well). There are three basic types of junctions, with the help of which we can characterize any pattern of arrows in the network.

1. ***A → B → C.*** This junction is the simplest example of a "chain," or of mediation. In science, one often thinks of *B* as the mechanism, or "mediator," that transmits the effect of *A* to *C*. A familiar example is Fire → Smoke → Alarm. Although we call them "fire alarms," they are really smoke alarms. The fire by itself does not set off an alarm, so there is no direct arrow from Fire to Alarm. Nor does the fire set off the alarm through any other variable, such as heat. It works only by releasing smoke molecules in the air. If we disable that link in the chain, for instance by sucking all the smoke molecules away with a fume hood, then there will be no alarm.

   This observation leads to an important conceptual point about chains: the mediator *B* "screens off" information about *A* from *C*, and vice versa. (This was first pointed out by Hans Reichenbach, a German-American philosopher of science.) For example, once we know the value of Smoke, learning about Fire does not give us any reason to raise or

advance in the development of Bayesian networks entailed finding ways to leverage sparseness in the network structure to achieve reasonable computation times.

BAYESIAN NETWORKS IN THE REAL WORLD

Bayesian networks are by now a mature technology, and you can buy off-the-shelf Bayesian network software from several companies. Bayesian networks are also embedded in many "smart" devices. To give you an idea of how they are used in real-world applications, let's return to the Bonaparte DNA-matching software with which we began this chapter.

The Netherlands Forensic Institute uses Bonaparte every day, mostly for missing-persons cases, criminal investigations, and immigration cases. (Applicants for asylum may have to take a DNA test to establish kinship to ~~must prove that they have fifteen~~ family members in the Netherlands.) However, the Bayesian network does its most impressive work after a massive disaster, such as the crash of Malaysia Airlines Flight 17.

Few, if any, of the victims of the plane crash could be identified by comparing DNA from the wreckage to DNA in a central database. The next best thing to do was to ask family members to provide DNA swabs and look for partial matches to the DNA of the victims. Conventional (non-Bayesian) methods can do this and have been instrumental in solving a number of cold cases in the Netherlands, the United States, and elsewhere. For example, a simple formula called the "Paternity Index" or the "Sibling Index" can estimate the likelihood that the unidentified DNA comes from the father or the brother of the person whose DNA was tested.

However, these indices are inherently limited because they work for only one specified relation and only for close relations. The idea behind Bonaparte is to make it possible to use DNA information from more distant relatives or from multiple relatives. Bonaparte does this by converting the pedigree of the family (see Figure 3.7) into a Bayesian network.

In Figure 3.8, we see how Bonaparte converts one small piece of a pedigree to a (causal) Bayesian network. The central problem is that the genotype of an individual, detected in a DNA test, contains a

to be independent, conditional on B, then we can safely conclude that the chain model is incompatible with the data and needs to be discarded (or repaired). Second, the graphical properties of the diagram dictate which causal models can be distinguished by data and which will forever remain indistinguishable, no matter how large the data. For example, we cannot distinguish the fork $A \leftarrow B \rightarrow C$ from the chain $A \rightarrow B \rightarrow C$ by data alone, because the two ~~diagrams~~ imply the same independence conditions.

Another convenient way of thinking about the causal model is in terms of hypothetical experiments. Each arrow can be thought of as a statement about the outcome of a hypothetical experiment. An arrow from A to C means that if we could wiggle only A, then we would expect to see a change in the probability of C. A missing arrow from A to C means that in the same experiment we would not see any change in C, once we held constant the parents of C (in other words, B in the example above). Note that the probabilistic expression "once we know the value of B" has given way to the causal expression "once we hold B constant," which implies that we are physically preventing B from varying and disabling the arrow from A to B.

The causal thinking that goes into the construction of the causal network will pay off, of course, in the type of questions the network can answer. Whereas a Bayesian network can only tell us how likely one event is, given that we observed another (rung-one information), causal diagrams can answer interventional and counterfactual questions. For example, the causal fork $A \leftarrow B \rightarrow C$ tells us in no uncertain terms that wiggling A would have no effect on C, no matter how intense the wiggle. On the other hand, a Bayesian network is not equipped to handle a "wiggle," or to tell the difference between seeing and doing, or indeed to distinguish a fork from a chain. In other words, both a chain and a fork would predict that observed changes in A are associated with changes in C, making no prediction about the effect of "wiggling" A.

Now we come to the second, and perhaps more important, impact of Bayesian networks on causal inference. The relationships that were discovered between the graphical structure of the

The same ambiguity plagues the third-variable definition. Should a confounder be a common cause of both X and Y or merely correlated with each? Today we can answer such questions by referring to the causal diagram and checking which variables produce a discrepancy between $P(Y \mid X)$ and $P(Y \mid do(X))$. Lacking a diagram or a *do*-operator, five generations of statisticians and health scientists had to struggle with surrogates, none of which were satisfactory. Considering that the drugs in your medicine cabinet may have been developed on the basis of a dubious definition of "confounders," you should be somewhat concerned.

Let's take a look at some of the surrogate definitions of confounding. These fall into two main categories, declarative and procedural. A typical (and wrong) declarative definition would be "A confounder is any variable that is correlated with both X and Y." On the other hand, a procedural definition would attempt to characterize a confounder in terms of a statistical test. This appeals to statisticians, who love any test that can be performed on the data directly without appealing to a model.

Here is a procedural definition that goes by the scary name of "noncollapsibility." It comes from a 1996 paper by the Norwegian epidemiologist Sven Hernberg: "Formally one can compare the crude relative risk and the relative risk resulting after adjustment for the potential confounder. A difference indicates confounding, and in that case one should use the adjusted risk estimate. If there is no or a negligible difference, confounding is not an issue and the crude estimate is to be preferred." In other words, if you suspect a confounder, try adjusting for it and try not adjusting for it. If there is a difference, it is a confounder, and you should trust the adjusted value. If there is no difference, you are off the hook. Hernberg was by no means the first person to advocate such an approach; it has misguided a century of epidemiologists, economists, and social scientists, and it still reigns in certain quarters of applied statistics. I have picked on Hernberg only because he was unusually explicit about it and because he wrote this in 1996, well after the Causal Revolution was already underway.

The most popular of the declarative definitions evolved over a period of time. Alfredo Morabia, author of *A History of*

the causal effect of X on Y. It is a disaster to control for Z if you are trying to find the causal effect of X on Y. If you look only at those individuals in the treatment and control groups for whom $Z = 0$, then you have completely blocked the effect of X, because it works by changing Z. So you will conclude that X has no effect on Y. This is exactly what Ezra Klein meant when he said, "Sometimes you end up controlling for the thing you're trying to measure."

In example (ii), Z is a proxy for the mediator M. Statisticians very often control for proxies when the actual causal variable can't be measured; for instance, party affiliation might be used as a proxy for political beliefs. Because Z isn't a perfect measure of M, some of the influence of X on Y might "leak through" if you control for Z. Nevertheless, controlling for Z is still a mistake. While the bias might be less than if you controlled for M, it is still there.

For this reason later statisticians, notably David Cox in his textbook *The ~~Design~~ Planning of Experiments* (1958), warned that you should only control for Z if you have a "strong prior reason" to believe that it is ~~not affected~~ "quite unaffected" by X. This ~~"strong prior reason" is nothing more or less than~~ "quite unaffected" condition is of course a causal assumption. He adds, "Such hypotheses may be perfectly in order, but the scientist should always be aware when they are being appealed to." Remember that it's 1958, in the midst of the great prohibition on causality. Cox is saying that you can go ahead and take a swig of causal moonshine when adjusting for confounders, but don't tell the preacher. A daring suggestion! I never fail to commend him for his bravery.

By 1980, Simpson's and Cox's conditions had been combined into the three-part test for confounding that I mentioned above. It is about as trustworthy as a canoe with only three leaks. Even though it does make a halfhearted appeal to causality in part (3), each of the first two parts can be shown to be both unnecessary and insufficient.

Greenland and Robins drew that conclusion in their landmark 1986 paper. The two took a completely new approach to confounding, which they called "exchangeability." They went back to the original idea that the control group ($X = 0$) should be comparable to the treatment group ($X = 1$). But they added a counterfactual twist. (Remember from Chapter 1 that counterfactuals are at rung

to assist in distinguishing confounders from deconfounders. She is the only person I know of who managed this feat. Later, in 2012, she collaborated on an updated version that analyzes the same examples with causal diagrams and verifies that all her conclusions from 1993 were correct.

In both of Weinberg's papers, the medical application was to estimate the effect of smoking (X) on miscarriages, or "spontaneous abortions" (Y). In Game 1, A represents an underlying abnormality that is induced by smoking; this is not an observable variable because we don't know what the abnormality is. B represents a history of previous miscarriages. It is very, very tempting for an epidemiologist to take previous miscarriages into account and adjust for them when estimating the probability of future miscarriages. But that is the wrong thing to do here! By doing so we are partially inactivating the mechanism through which smoking acts, and we will thus underestimate the true effect of smoking.

Game 2 is a more complicated version where there are two different smoking variables: X represents whether the mother smokes now (at the beginning of the second pregnancy), while A represents whether she smoked during the first pregnancy. B and E are underlying abnormalities caused by smoking, which are unobservable, and D represents other physiological causes of those abnormalities. Note that this diagram allows for the fact that the mother could have changed her smoking behavior between pregnancies, but the other physiological causes would not change. Again, many epidemiologists would adjust for prior miscarriages (C), but this is a bad idea unless you also adjust for smoking behavior in the first pregnancy (A).

Games 4 and 5 come from a paper published in 2014 by Andrew Forbes, a biostatistician at Monash University in Australia, along with several collaborators. He is interested in the effect of smoking on adult asthma. In Game 4, X represents an individual's smoking behavior, and Y represents whether the person has asthma as an adult. B represents childhood asthma, which is a collider because it is affected by both A, parental smoking, and C, an underlying (and unobservable) predisposition toward asthma. In Game 5 the variables have the same meanings, but Forbes added two arrows

for greater realism. (Game 4 was only meant to introduce the *M*-graph.)

In fact, the full model in ~~Forbes'~~ their paper has a few more variables and looks like the diagram in Figure 4.7. Note that Game 5 is embedded in this model in the sense that the variables *A*, *B*, *C*, *X*, and *Y* have exactly the same relationships. So we can transfer our conclusions over and conclude that we have to control for *A* and *B* or for *C*; but *C* is an unobservable and therefore uncontrollable variable. In addition we have four new confounding variables: *D* = parental asthma, *E* = chronic bronchitis, *F* = sex, and *G* = socioeconomic status. The reader might enjoy figuring out that we must control for *E*, *F*, and *G*, but there is no need to control for *D*. So a sufficient set of variables for deconfounding is *A*, *B*, *E*, *F*, and *G*.

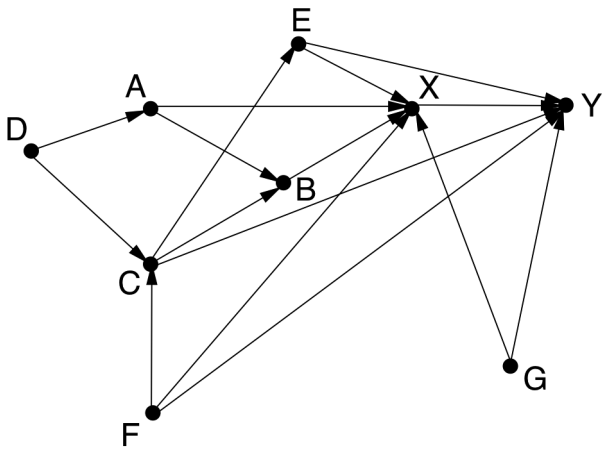

FIGURE 4.7. Andrew Forbes's ~~'s~~ and Elizabeth Williamson's model of smoking (*X*) and asthma (*Y*).

In the end, Forbes and Williamson found that smoking had a small and statistically insignificant association with adult asthma in the raw data, and the effect became even smaller and more insignificant after adjusting for the confounders. The null result should not detract, however, from the fact that ~~his~~ their paper is a model for the "skillful interrogation of Nature."

it might be affected by one of those other causes. If we find out that the mother is a smoker, this explains away the low weight and consequently reduces the likelihood of a serious birth defect. But if the mother does not smoke, we have stronger evidence that the cause of the low birth weight is a birth defect, and the baby's prognosis becomes worse.

As before, a causal diagram makes everything clearer. When we incorporate the new assumptions, the causal diagram looks like Figure 5.4. We can see that the birth-weight paradox is a perfect example of collider bias. The collider is Birth Weight. By looking only at babies with low birth weight, we are conditioning on that collider. This opens up an additional ~~a back door~~ path between Smoking and Mortality that goes Smoking → Birth Weight ← Birth Defect → Mortality. This path is noncausal because one of the arrows goes the wrong way. Nevertheless, it induces a spurious correlation between Smoking and Mortality and biases our estimate of the actual (direct) causal effect, Smoking → Mortality. In fact, it biases the estimate to such a large extent that smoking actually appears beneficial.

The beauty of causal diagrams is that they make the source of bias obvious. Lacking such diagrams, epidemiologists argued about the paradox for forty years. In fact, they are still discussing it: the October 2014 issue of the *International Journal of Epidemiology* contains several articles on this topic. One of them, by Tyler VanderWeele of Harvard, nails the explanation perfectly and contains a diagram just like the one below.

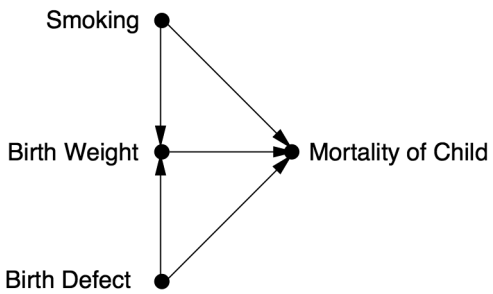

FIGURE 5.4. Causal diagram for the birth-weight paradox.

patient chooses to take Drug *D*. In the study, women clearly had a preference for taking Drug *D* and men preferred not to. Thus Gender is a confounder of Drug and Heart Attack. For an unbiased estimate of the effect of Drug on Heart Attack, we must adjust for the confounder. We can do that by looking at the data for men and women separately, then taking the average:

- For women, the rate of heart attacks was 5 percent without Drug *D* and 7.5 percent with Drug *D*.
- For men, the rate of heart attacks was 30 percent without Drug *D* and 40 percent with.
- Taking the average (because men and women are equally frequent in the general population), the rate of heart attacks without Drug *D* is 17.5 percent (the average of 5 and 30), and the rate with Drug *D* is 23.75 percent (the average of 7.5 and 40).

This is the clear and unambiguous answer we were looking for. Drug *D* isn't BBG, it's BBB: bad for women, bad for women, and bad for people.

I don't want you to get the impression from this example that aggregating the data is always wrong or that partitioning the data is always right. It depends on the process that generated the data. In the Monty Hall paradox, we saw that changing the rules of the game also changed the conclusion. The same principle works here. I'll use a different story to demonstrate when pooling the data would be appropriate. Even though the data will be precisely the same, the role of the "lurking third variable" will differ and so will the conclusion.

Let's begin with the assumption that blood pressure is known to be a possible cause of heart attack, and Drug *B* is supposed to reduce blood pressure. Naturally, the Drug *B* researchers wanted to see if it might also reduce heart attack risk, so they measured their patients' blood pressure after treatment, as well as whether they had a heart attack.

Table 6.6 shows the data from the study of Drug *B*. It should look amazingly familiar: the numbers are the same as in Table

blood pressure because the blood pressure measurement comes after the patient takes the drug, but we should stratify the data in the case of gender because it is determined before the patient takes the drug. While this rule will work in a great many cases, it is not foolproof. A simple case is that of M-bias (Game 4 in Chapter 4). Here B can precede X; yet we should still not condition on B, because that would violate the back-door criterion. We should consult the causal structure of the story, not the temporal information.

Finally, you might wonder if Simpson's paradox occurs in the real world. The answer is yes. It is certainly not common enough for statisticians to encounter on a daily basis, but nor is it completely unknown, and it probably happens more often than journal articles report. Here are two documented cases:

- In an observational study published in 1996, open surgery to remove kidney stones had a better success rate than endoscopic surgery for small kidney stones. It also had a better success rate for large kidney stones. However, it had a lower success rate overall. Just as in our first example, this was a case where the choice of treatment was related to the severity of the patients' case: larger stones were more likely to lead to open surgery and also had a worse prognosis.
- In a study of thyroid disease published in 1995, smokers had a higher survival rate (76 percent) over twenty years than nonsmokers (69 percent). However, the nonsmokers had a better survival rate in six out of seven age groups, and the difference was minimal in the seventh. Age was clearly a confounder of Smoking and Survival: the average smoker was younger than the average nonsmoker (perhaps because the older smokers had already died). Stratifying the data by age, we conclude that smoking has a negative impact on survival.

Because Simpson's paradox has been so poorly understood, some statisticians take precautions to avoid it. All too often, these methods avoid the symptom, Simpson's reversal, without doing anything about the disease, confounding. Instead of suppressing

figure. In Figure 6.7, we can see that boys gain more weight than girls in every stratum (every vertical cross section). Yet it's equally obvious that both boys and girls gained nothing overall. How can that be? Is not the overall gain just an average of the stratum-specific gains?

Now that we are experienced pros at the fine points of Simpson's paradox and the sure-thing principle, we know what is wrong with that argument. The sure-thing principle works only in cases where the relative proportion of each subpopulation (each weight class) does not change from group to group. Yet, in Lord's case, the "treatment" (gender) very strongly affects the percentage of students in each weight class.

So we can't rely on the sure-thing principle, and that brings us back to square one. Who is right? Is there or isn't there a difference in the average weight gains between boys and girls when proper allowance is made for differences in the initial weight between the sexes? Lord's conclusion is very pessimistic: "The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data." Lord's pessimism spread beyond statistics and has led to a rich and quite pessimistic literature in epidemiology and biostatistics on how to compare groups that differ in "baseline" statistics.

I will show now why Lord's pessimism is unjustified. The dietitian's question can be answered in a rigorous way, and as usual the starting point is to draw a causal diagram, as in Figure 6.8. In this diagram, we see that Sex (S) is a cause of initial weight (W_I) and final weight (W_F). Also, W_I affects W_F independently of gender, because students of either gender who weigh more at the beginning of the year tend to weigh more at the end of the year, as shown by the scatter plots in Figure 6.7. All these causal assumptions are commonsensical; I would not expect Lord to disagree with them.

The variable of interest to Lord is the weight gain, shown as Y in this diagram. Note that Y is related to W_I and W_F in a purely mathematical, deterministic way: $Y = W_F - W_I$. ~~This means that the correlations between Y and W_I (or Y and W_F) are equal to −1 (or 1), and~~ I have shown this information on the diagram with the coefficients –1 and +1.

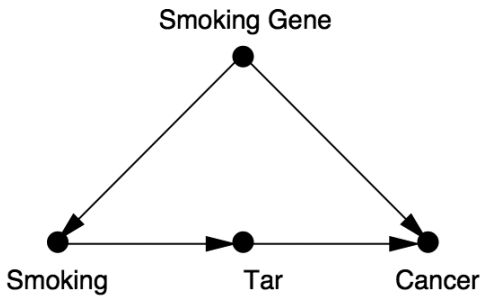

FIGURE 7.1. Hypothetical causal diagram for smoking and cancer, suitable for front-door adjustment.

no direct arrow points from Smoking to Cancer, and there are no other indirect pathways.

Suppose we are doing an observational study and have collected data on Smoking, Tar, and Cancer for each of the participants. Unfortunately, we cannot collect data on the Smoking Gene because we do not know whether such a gene exists. Lacking data on the confounding variable, we cannot block the back-door path Smoking ← Smoking Gene → Cancer. Thus we cannot use back-door adjustment to control for the effect of the confounder.

So we must look for another way. Instead of going in the back door, we can go in the front door! In this case, the front door is the direct causal path Smoking → Tar → Cancer, for which we do have data on all three variables. Intuitively, the reasoning is as follows. First, we can estimate the average causal effect of Smoking on Tar, because there is no unblocked back-door path from Smoking to ~~Cancer,~~ Tar, as the Smoking ← Smoking Gene → Cancer ← Tar path is already blocked by the collider at Cancer. Because it is blocked already, we don't even need back-door adjustment. We can simply observe $P(tar \mid smoking)$ and $P(tar \mid no\ smoking)$, and the difference between them will be the average causal effect of Smoking on Tar.

Likewise, the diagram allows us to estimate the average causal effect of Tar on Cancer. To do this we can block the back-door path from Tar to Cancer, Tar ← Smoking ← Smoking Gene → Cancer,

by adjusting for Smoking. Our lessons from Chapter 4 come in handy: we only need data on a sufficient set of deconfounders (i.e., Smoking). Then the back-door adjustment formula will give us $P(cancer \mid do(tar))$ and $P(cancer \mid do(no\ tar))$. The difference between these is the average causal effect of Tar on Cancer.

Now we know the average increase in the likelihood of tar deposits due to smoking and the average increase of cancer due to tar deposits. Can we combine these somehow to obtain the average increase in cancer due to smoking? Yes, we can. The reasoning goes as follows. Cancer can come about in two ways: in the presence of Tar or in the absence of Tar. If we force a person to smoke, then the probabilities of these two states are $P(tar \mid do(smoking))$ and $P(no\ tar \mid do(\cancel{no}\ smoking))$, respectively. If a Tar state evolves, the likelihood of causing Cancer is $P(cancer \mid do(tar))$. If, on the other hand, a No-Tar state evolves, then it would result in a Cancer likelihood of $P(cancer \mid do(no\ tar))$. We can weight the two scenarios by their respective probabilities under $do(smoking)$ and in this way compute the total probability of cancer due to smoking. The same argument holds if we prevent a person from smoking, $do(no\ smoking)$. The difference between the two gives us the average causal effect on cancer of smoking versus not smoking.

As I have just explained, we can estimate each of the *do*-probabilities discussed from the data. That is, we can write them mathematically in terms of probabilities that do not involve the *do*-operator. In this way, mathematics does for us what ten years of debate and congressional testimony could not: quantify the causal effect of smoking on cancer—provided our assumptions hold, of course.

The process I have just described, expressing $P(cancer \mid do(smoking))$ in terms of *do*-free probabilities, is called the front-door adjustment. It differs from the back-door adjustment in that we adjust for two variables (Smoking and Tar) instead of one, and these variables lie on the front-door path from Smoking to Cancer rather than the back-door path. For those readers who "speak mathematics," I can't resist showing you the formula (Equation 7.1), which cannot be found in ordinary statistics textbooks. Here X stands for Smoking, Y stands for Cancer, Z stands for Tar, and

U (which is conspicuously absent from the formula) stands for the unobservable variable, the Smoking Gene.

$$P(Y \mid do(X)) = \Sigma_z P(Z = z \mid X) \, \Sigma_x P(Y \mid X = x, Z = z) \, P(X = x) \tag{7.1}$$

Readers with an appetite for mathematics might find it interesting to compare this to the formula for the back-door adjustment, which looks like Equation 7.2.

$$P(Y \mid do(X)) = \Sigma_z P(Y \mid X, Z = z) \, P(Z = z) \tag{7.2}$$

Even for readers who do not speak mathematics, we can make several interesting points about Equation 7.1. First and most important, you don't see U (the Smoking Gene) anywhere. This was the whole point. We have successfully deconfounded U even without possessing any data on it. Any statistician of Fisher's generation would have seen this as an utter miracle. Second, way back in the Introduction I talked about an estimand as a recipe for computing the quantity of interest in a query. Equations 7.1 and 7.2 are the most complicated and interesting estimands that I will show you in this book. The left-hand side represents the query "What is the effect of X on Y?" The right-hand side is the estimand, a recipe for answering the query. Note that the estimand contains no *do*'s, only *see*'s, represented by the vertical bars, and this means it can be estimated from data.

At this point, I'm sure that some readers are wondering how close this fictional scenario is to reality. Could the smoking-cancer controversy have been resolved by one observational study and one causal diagram? If we assume that Figure 7.1 accurately reflects the causal mechanism for cancer, the answer is absolutely yes. However, we now need to discuss whether our assumptions are valid in the real world.

David Freedman, a longtime friend and a Berkeley statistician, took me to task over this issue. He argued that the model in Figure 7.1 is unrealistic in three ways. First, if there is a smoking gene, it might also affect how the body gets rid of foreign matter in the

benchmarks by hundreds or thousands of dollars. This is exactly what you would expect to see if there is an unobserved confounder, such as Motivation. The back-door criterion cannot adjust for it.

On the other hand, the front-door estimates succeeded in removing almost all of the Motivation effect. For males, the front-door estimates were well within the experimental error of the randomized controlled trial, even with the small positive bias that Glynn and Kashin predicted. For females, the results were even better: The front-door estimates matched the experimental benchmark almost perfectly, with no apparent bias. Glynn and Kashin's work gives both empirical and methodological proof that as long as the effect of C on M (in Figure 7.2) is weak, front-door adjustment can give a reasonably good estimate of the effect of X on Y. It is much better than not controlling for C.

Glynn and Kashin's results show why the front-door adjustment is such a powerful tool: it allows us to control for confounders that we cannot observe (like Motivation), including those that we can't even name. RCTs are considered the "gold standard" of causal effect estimation for exactly the same reason. Because front-door estimates do the same thing, with the additional virtue of observing people's behavior in their own natural habitat instead of a laboratory, I would not be surprised if this method eventually becomes a ~~serious competitor~~ useful alternative to randomized controlled trials.

THE *DO*-CALCULUS, OR MIND OVER MATTER

In both the front- and back-door adjustment formulas, the ultimate goal is to calculate the effect of an intervention, $P(Y \mid do(X))$, in terms of data such as $P(Y \mid X, A, B, Z, \ldots)$ that do not involve a *do*-operator. If we are completely successful at eliminating the *do*'s, then we can use observational data to estimate the causal effect, allowing us to leap from rung one to rung two of the Ladder of Causation.

The fact that we were successful in these two cases (front- and back-door) immediately raises the question of whether there are other doors through which we can eliminate all the *do*'s. Thinking more generally, we can ask whether there is some way to decide in

DO-CALCULUS AT WORK

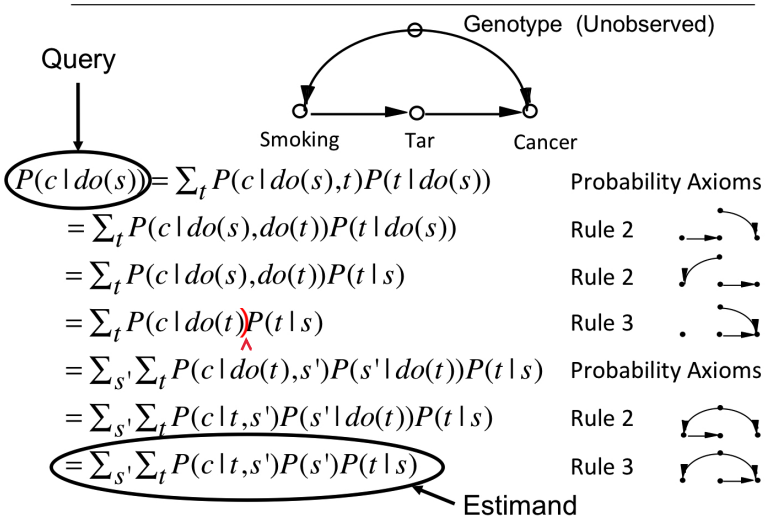

FIGURE 7.4. Derivation of the front-door adjustment formula from the rules of *do*-calculus.

for confounders. I believed no one could do this without the *do*-calculus, so I presented it as a challenge in a statistics seminar at Berkeley in 1993 and even offered a $100 prize to anyone who could solve it. Paul Holland, who attended the seminar, wrote that he had assigned the problem as a class project and would send me the solution when ripe. (Colleagues tell me that he eventually presented a long solution at a conference in 1995, and I may owe him $100 if I could only find his proof.) Economists James Heckman and Rodrigo Pinto made the next attempt to prove the front-door formula using "standard tools" in 2015. They succeeded, albeit at the cost of eight pages of hard labor.

In a restaurant the evening before the talk, I had written the proof (very much like the one in Figure 7.4) on a napkin for David Freedman. He wrote me later to say that he had lost the napkin. He could not reconstruct the argument and asked if I had kept a copy. The next day, Jamie Robins wrote to me from Harvard, saying that he had heard about the "napkin problem" from Freedman, and he straightaway offered to fly to California to check the

determine how many lives would have been saved by purifying the water supply.

Here's how the trick works. For simplicity we'll go back to the names Z, X, Y, and U for our variables and redraw Figure 7.8 as seen in Figure 7.9. I have included path coefficients (a, b, c, d) to represent the strength of the causal effects. This means we are assuming that the variables are numerical and the functions relating them are linear. Remember that the path coefficient a means that an intervention to increase Z by one standard unit will cause X to increase by a standard units. (I will omit the technical details of what the "standard units" are.)

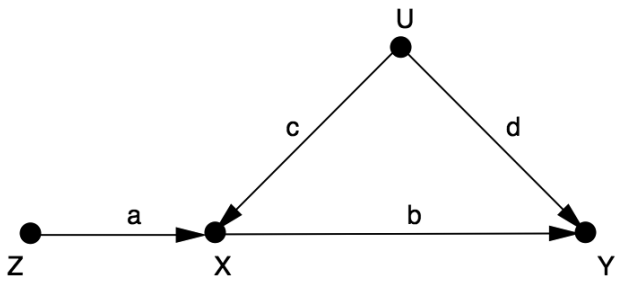

FIGURE 7.9. General setup for instrumental variables.

Because Z and X are unconfounded, the causal effect of Z on X (that is, a) can be estimated from the slope r_{XZ} of the regression line of X on Z. Likewise, the variables Z and Y are unconfounded, because the path $Z \rightarrow X \leftarrow U \rightarrow Y$ is blocked by the collider at X. So the slope of the regression line of Y on Z (r_{YZ}) will equal the causal effect on the direct path $Z \rightarrow X \rightarrow Y$, which is the product of the path coefficients: ab. Thus we have two equations: $ab = r_{YZ}$ and $a = r_{XZ}$. If we divide the first equation by the second, we get the causal effect of X on Y: $b = r_{YZ}/r_{XZ}$.

In this way, instrumental variables allow us to perform the same kind of magic trick that we did with front-door adjustment: we have found the effect of X on Y even without being able to control

for, or collect data on, the confounder, U. We can therefore provide decision makers with a conclusive argument that they should move their water supply—even if those decision makers still believe in the miasma theory. Also notice that we have gotten information on the second rung of the Ladder of Causation (b) from information about the first rung (the correlations, r_{YZ} and r_{XZ}). We were able to do this because the assumptions embodied in the path diagram are causal in nature, especially the crucial assumption that there is no arrow between U and Z. If the causal diagram were different—for example, if Z were a confounder of X and Y—the formula $b = r_{YZ}/r_{XZ}$ would not correctly estimate the causal effect of X on Y. In fact, these two models cannot be told apart by any statistical method, regardless of how big the data.

Instrumental variables were known before the Causal Revolution, but causal diagrams have brought new clarity to how they work. Indeed, Snow was using an instrumental variable implicitly, although he did not have a quantitative formula. Sewall Wright certainly understood this use of path diagrams; the formula $b = r_{YZ}/r_{XZ}$ can be derived directly from his method of path coefficients. And it seems that the first person other than Sewall Wright to use instrumental variables in a deliberate way was . . . Sewall Wright's father, Philip!

Recall that Philip Wright was an economist who worked at what later became the Brookings Institution. He was interested in predicting how the output of a commodity would change if a tariff were imposed, which would raise the price and therefore, in theory, encourage production. In economic terms, he wanted to know the elasticity of supply.

In 1928 Wright wrote a long monograph dedicated to computing the elasticity of supply for flaxseed oil. In a remarkable appendix, he analyzed the problem using a path diagram. This was a brave thing to do: remember that no economist had ever seen or heard of such a thing before. (In fact, he hedged his bets and verified his calculations using more traditional methods.)

Figure 7.10 shows a somewhat simplified version of Wright's diagram. Unlike most diagrams in this book, this one has "two-way"

me, this historical detective work makes the story more beautiful. It shows that Philip took the trouble to understand his son's theory and articulate it in his own language.

Now let's move forward from the 1850s and 1920s to look at a present-day example of instrumental variables in action, one of literally dozens I could have chosen.

GOOD AND BAD CHOLESTEROL

Do you remember when your family doctor first started talking to you about "good" and "bad" cholesterol? It may have happened in the 1990s, when drugs that lowered blood levels of "bad" cholesterol, low-density lipoprotein (LDL), first came on the market. These drugs, called statins, have turned into multibillion-dollar revenue generators for pharmaceutical companies.

An early ~~The first~~ cholesterol-modifying drug subjected to a randomized controlled trial was cholestyramine. The Coronary Primary Prevention Trial, begun in 1973 and concluded in 1984, showed a 12.6 percent reduction in cholesterol among men given the drug cholestyramine and a 19 percent reduction in the risk of heart attack.

Because this was a randomized controlled trial, you might think we wouldn't need any of the methods in this chapter, because they are specifically designed to replace RCTs in situations where you only have observational data. But that is not true. This trial, like many RCTs, faced the problem of noncompliance, when subjects randomized to receive a drug don't actually take it. This will reduce the apparent effectiveness of the drug, so we may want to adjust the results to account for the noncompliers. But as always, confounding rears its ugly head. If the noncompliers are different from the compliers in some relevant way (maybe they are sicker to start with?), we cannot predict how they would have responded had they adhered to instructions.

In this situation, we have a causal diagram that looks like Figure 7.11. The variable Assigned (Z) will take the value 1 if the patient is randomly assigned to receive the drug and 0 if he is randomly assigned a placebo. The variable Received will be 1 if the patient actually took the drug and 0 otherwise. For convenience, we'll also

While Thucydides and Abraham probed counterfactuals through individual cases, the Greek philosopher Aristotle investigated more generic aspects of causation. In his typically systematic style, Aristotle set up a whole taxonomy of causation, including "material causes," "formal causes," "efficient causes," and "final causes." For example, the material cause of the shape of a statue is the bronze from which it is cast and its properties; we could not make the same statue out of Silly Putty. However, Aristotle nowhere makes a statement about causation as a counterfactual, so his ingenious classification lacks the simple clarity of Thucydides's account of the cause of the tsunami.

To find a philosopher who placed counterfactuals at the heart of causality, we have to move ahead to David Hume, the Scottish philosopher and contemporary of Thomas Bayes. Hume rejected Aristotle's classification scheme and insisted on a single definition of causation. But he found this definition quite elusive and was in fact torn between two different definitions. Later these would turn into two incompatible ideologies, which ironically could both cite Hume as their source!

In his *Treatise of Human Nature* (Figure 8.1), Hume denies that any two objects have innate qualities or "powers" that make one a cause and the other an effect. In his view, the cause-effect relationship is entirely a product of our own memory and experience. "Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*," he writes. "We likewise call to mind their constant conjunction in all past instances. Without any further ceremony, we call the one *cause* and the other *effect*, and infer the existence of the one from the other." This is now known as the "regularity" definition of causation.

The passage is breathtaking in its chutzpah. Hume is cutting off the second and third rungs of the Ladder of Causation and saying that the first rung, observation, is all that we need. Once we observe flame and heat together a sufficient number of times (and note that flame has temporal precedence), we agree to call flame the cause of heat. ~~Like most twentieth-century statisticians,~~ Not unlike Karl Pearson 200 years later, Hume in 1739 seems happy to consider causation as merely a species of correlation.

This definition is perfectly legitimate for someone in possession of a probability function over counterfactuals. But how is a biologist or economist with only scientific knowledge for guidance supposed to assess whether this is true or not? More concretely, how is a scientist to assess whether ignorability holds in any of the examples discussed in this book?

To understand the difficulty, let us attempt to apply this explanation to our example. To determine if *ED* is ignorable (conditional on *EX*), we are supposed to judge whether employees who would have one potential salary, say $S_1 = s$, are just as likely to have one level of education as the employees who would have a different potential salary, say $S_1 = s'$. If you think that this sounds circular, I can only agree with you! We want to determine Alice's potential salary, and even before we start—even before we get a hint about the answer—we are supposed to speculate on whether the result is dependent or independent of *ED*, in every stratum of *EX*. It is quite a cognitive nightmare.

As it turns out, *ED* in our example is not ignorable with respect to *S*, conditional on *EX*, and this is why the matching approach (setting Bert and Caroline equal) would yield the wrong answer for their potential salaries. In fact, their estimates should differ by an amount $S_1(\text{Bert}) - S_1(\text{Caroline}) = -\$9{,}500$. The reader should be able to show this from the numbers in Table 8.1 and the three-step procedure.) I will now show that with the help of a causal diagram, a student could see immediately that *ED* is not ignorable and would not attempt matching here. Lacking a diagram, a student would be tempted to assume that ignorability holds by default and would fall into this trap. (This is not a speculation. I borrowed the idea for this example from an article in *Harvard Law Review* where the story was essentially the same as in Figure 8.3 and the author did use matching.)

Here is how we can use a causal diagram to test for (conditional) ignorability. To determine if *X* is ignorable relative to outcome *Y*, conditional on a set *Z* of matching variables, we need only test to see if *Z* blocks all the back-door paths between *X* and *Y* and no member of *Z* is a descendant of *X*. It is as simple as that! In our example, the proposed matching variable (Experience) blocks

Joe is legally responsible for her death even though he did not light the fire.

How can we express necessary or but-for causes in terms of potential outcomes? If we let the outcome Y be "Judy's death" (with $Y = 0$ if Judy lives and $Y = 1$ if Judy dies) and the treatment X be "Joe's blocking the fire escape" (with $X = 0$ if he does not block it and $X = 1$ if he does), then we are instructed to ask the following question:

> Given that we know the fire escape was blocked ($X = 1$) and Judy died ($Y = 1$), what is the probability that Judy would have lived ($Y = 0$) if X had been 0?

Symbolically, the probability we want to evaluate is $P(Y_{X=0} = 0 \mid X = 1, Y = 1)$. Because this expression is rather cumbersome, I will later abbreviate it as "*PN*," the *probability of necessity* (i.e., the probability that $X = 1$ is a necessary or but-for cause of $Y = 1$).

italicize

Note that the probability of necessity involves a contrast between two different worlds: the actual world where $X = 1$ and the counterfactual world where $X = 0$ (expressed by the subscript $X = 0$). In fact, hindsight (knowing what happened in the actual world) is a critical distinction between counterfactuals (rung three of the Ladder of Causation) and interventions (rung two). Without hindsight, there is no difference between $P(Y_{X=0} = 0)$ and $P(Y = 0 \mid do(X = 0))$. Both express the probability that, under normal conditions, Judy will be alive if we ensure that the exit is not blocked; they do not mention the fire, Judy's death, or the blocked exit. But hindsight may change our estimate of the probabilities. Suppose we observe that $X = 1$ and $Y = 1$ (hindsight). Then $P(Y_{X=0} = 0 \mid X = 1, Y = 1)$ is not the same as $P(Y_{X=0} = 0 \mid X = 1)$. Knowing that Judy died ($Y = 1$) gives us information on the circumstances that we would not get just by knowing that the door was blocked ($X = 1$). For one thing, it is evidence of the strength of the fire.

In fact, it can be shown that there is no way to capture $P(Y_{X=0} = 0 \mid X = 1, Y = 1)$ in a *do*-expression. While this may seem like a rather arcane point, it does give mathematical proof

As a disciple of Terman, Burks must have been disappointed to see such a small effect. (In fact, her estimates have held up quite well over time.) So she questioned the then accepted method of analysis, which was to control for Social Status. "The true measure of contribution of a cause to an effect is mutilated," she wrote, "if we have rendered constant variables which may *in part or in whole be caused by either of the two factors whose true relationship is to be measured, or by still other unmeasured remote causes which also affect either of the two isolated factors*" (emphasis in the original). In other words, if you are interested in the total effect of Parental Intelligence on Child's Intelligence, you should not adjust for (render constant) any variable on the pathway between them.

But Burks didn't stop there. Her italicized criterion, translated into modern language, reads that a bias will be introduced if we condition on variables that are (a) effects of either Parental Intelligence or Child's Intelligence, or (b) effects of unmeasured causes of either Parental Intelligence or Child's Intelligence (such as *X* in Figure 9.2).

These criteria were far ahead of their time and unlike anything that Sewall Wright had written. In fact, criterion (b) is one of the earliest examples ever of collider bias. If we look at Figure 9.2, we see that Social Status is a collider (Parental Intelligence → Social Status ← *X*). Therefore, controlling for Social Status opens the ~~back door~~ path Parental Intelligence → Social Status ← *X* → Child's Intelligence. Any resulting estimate of the indirect and direct effects will be biased. Because statisticians before (and after) Burks did not think in terms of arrows and diagrams, they were totally immersed in the myth that, while simple correlation has no causal implications, controlled correlation (or partial regression coefficients, see p. 222) is a step in the direction of causal explanation.

Burks was not the first person to discover the collider effect, but one can argue that she was the first to characterize it generally in graphical terms. Her criterion (b) applies perfectly to the examples of M-bias in Chapter 4. Hers is the first warning ever against conditioning on a pretreatment factor, a habit deemed safe by all twentieth-century statisticians and oddly still considered safe by some.

for confounders between mediator and outcome. Yet those who eschew the language of diagrams (some economists still do) complain and confess that it is a torture to explain what this warning means.

Thankfully, the problem that Kruskal once called "perhaps insoluble" was solved two decades ago. I have this strange feeling that Kruskal would have enjoyed the solution, and in my fantasy I imagine showing him the power of the *do*-calculus and the algorithmization of counterfactuals. Unfortunately, he retired in 1990, just when the rules of *do*-calculus were being shaped, and he died in 2005.

I'm sure that some readers are wondering: What finally happened in the Berkeley case? The answer is, nothing. Hammel and Bickel were convinced that Berkeley had nothing to worry about, and indeed no lawsuits or federal investigations ever materialized. The data hinted at reverse discrimination against males, and in fact there was explicit evidence of this: "In most of the cases involving favored status for women it appears that the admissions committees were seeking to overcome long-established shortages of women in their fields," Bickel wrote. Just three years later, a lawsuit over affirmative action on another campus of the University of California went all the way to the Supreme Court. Had the Supreme Court struck down affirmative action, such "favored status for women" might have become illegal. However, the Supreme Court upheld affirmative action, and the Berkeley case became a historical footnote.

A wise man leaves the final word not with the Supreme Court but with his wife. Why did mine have such a strong intuitive conviction that it is utterly impossible for a school to discriminate while each of its departments acts fairly? It is a theorem of causal calculus similar to the sure-thing principle. The sure-thing principle, as Leonard ~~Jimmie~~ Savage originally stated it, pertains to total effects, while this theorem holds for direct effects. The very definition of a direct effect on a global level relies on aggregating direct effects in the subpopulations.

To put it succinctly, local fairness everywhere implies global fairness. My wife was right.

it can have a blueprint summary of its major software components. Other components can then reason about that blueprint and mimic a state of self-awareness.

To create the perception of agency, we must also equip this software package with a memory to record past activations, to which it can refer when asked, "Why did you do that?" Actions that pass certain patterns of path activation will receive reasoned explanations, such as "Because the alternative proved less attractive." Others will end up with evasive and useless answers, such as "I wish I knew why" or "Because that's the way you programmed me."

In summary, I believe that the software package that can give a thinking machine the benefits of agency would consist of at least three parts: a causal model of the world; a causal model of its own software, however superficial; and a memory that records how intents in its mind correspond to events in the outside world.

This may even be how our own causal education as infants begins. We may have something like an "intention generator" in our minds, which tells us that we are supposed to take action $X = x$. But children love to experiment—to defy their parents', their teachers', even their own initial intentions—and to do something different, just for fun. Fully aware that we are supposed to do $X = x$, we playfully do $X = x'$ instead. We watch what happens, repeat the process, and keep a record of how good our intention generator is. Finally, when we start to adjust our own software, that is when we begin to take moral responsibility for our actions. This responsibility may be an illusion at the level of neural activation but not at the level of self-awareness software.

Encouraged by these possibilities, I believe that strong AI with causal understanding and agency capabilities is a realizable promise, and this raises the question that science fiction writers have been asking since the 1950s: Should we be worried? Is strong AI a Pandora's box that we should not open?

Recently public figures like Elon Musk and Stephen Hawking have gone on record saying that we should be worried. On Twitter, Musk said that AIs were "potentially more dangerous than nukes." In 2015, John Brockman's website Edge.org posed as its annual question, that year asking, "What do you think about machines

Blai Bonet, Carlo Brito, Avin Chen, Bryant Chen, David Chickering, Adnan Darwiche, Rina Dechter, Andrew Forney, David Galles, Hector Geffner, Dan Geiger, Moises Goldszmidt, David Heckerman, Mark Hopkins, Jin Kim, Manabu Kuroki, Trent Kyono, Karthika Mohan, Azaria Paz, George Rebane, Ilya Shpitser, Jin Tian, ~~and~~ Thomas Verma, and Ingrid Zukerman

Funding agencies receive ritualized thanks in scholarly publications but far too little real credit, considering their crucial role in recognizing seeds of ideas before they become fashionable. I must acknowledge the steady and unfailing support of the National Science Foundation and the Office of Naval Research, through the Machine Learning and Intelligence program headed by Behzad Kamgar-Parsi.

Dana and I would like to thank our agent, John Brockman, who gave us timely encouragement and the benefit of his professional expertise. Our editor at Basic Books, TJ Kelleher, asked us just the right questions and persuaded Basic Books that a story this ambitious could not be told in 200 pages. Our illustrators, Maayan Harel and Dakota Harr, managed to cope with our sometimes conflicting instructions and brought abstract subjects to life with humor and beauty. Kaoru Mulvihill at UCLA deserves much credit for proofing several versions of the manuscript and illustrating the hordes of graphs and diagrams.

Dana will be forever grateful to John Wilkes, who founded the Science Communication Program at UC Santa Cruz, which is still going strong and is the best possible route into a career as a science writer. Dana would also like to thank his wife, Kay, who encouraged him to pursue his childhood dream of being a writer, even when it meant pulling up stakes, crossing the country, and starting over.

Finally, my deepest debt is owed to my family, for their patience, understanding, and support. Especially to my wife, Ruth, my moral compass, for her endless love and wisdom. To my late son, Danny, for showing me the silent audacity of truth. To my daughters Tamara and Michelle for trusting my perennial promise that the book will eventually be done. And to my grandchildren, Leora, Tori, Adam, Ari and Evan, for giving a purpose to my long journeys and for always dissolving my "why" questions away.

Lilienfeld, ~~A.~~ D. (2007). Abe and Yak: The interactions of Abraham M. Lilienfeld and Jacob Yerushalmy in the development of modern epidemiology (1945–1973). *Epidemiology* 18: 507–514.

Morabia, A. (2013). Hume, Mill, Hill, and the sui generis epidemiologic approach to causal inference. *American Journal of Epidemiology* 178: 1526–1532.

Parascandola, M. (2004). Two approaches to etiology: The debate over smoking and lung cancer in the 1950s. *Endeavour* 28: 81–86.

Proctor, R. (2012a). *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition*. University of California Press, Berkeley, CA.

Proctor, R. (2012b). The history of the discovery of the cigarette–lung cancer link: Evidentiary traditions, corporate denial, and global toll. *Tobacco Control* 21: 87–91.

Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Company, New York, NY.

Stolley, P. (1991). When genius errs: R. A. Fisher and the lung cancer controversy. *American Journal of Epidemiology* 133: 416–425.

US Department of Health and Human Services (USDHHS). (2014). The health consequences of smoking—50 years of progress: A report of the surgeon general. USDHHS and Centers for Disease Control and Prevention, Atlanta, GA.

VanderWeele, T. (2014). Commentary: Resolutions of the birthweight paradox: Competing explanations and analytical insights. *International Journal of Epidemiology* 43: 1368–1373.

Wilcox, A. (2001). On the importance—and the unimportance—of birthweight. *International Journal of Epidemiology* 30: 1233–1241.

Wilcox, A. (2006). The perils of birth weight—A lesson from directed acyclic graphs. *American Journal of Epidemiology* 164: 1121–1123.

Wingo, P. (2003). Long-term trends in cancer mortality in the United States, 1930–1998. *Cancer* 97: 3133–3275.

CHAPTER 6. PARADOXES GALORE!

Annotated Bibliography

The Monty Hall paradox appears in many introductory books on probability theory (e.g., Grinstead and Snell, 1998, p. 136; Lindley,

2014, p. 201). The equivalent "three prisoners dilemma" was used to demonstrate the inadequacy of non-Bayesian approaches in Pearl (1988, pp. 58–62).

Tierney (July 21, 1991) and Crockett (2015) tell the amazing story of vos Savant's column on the Monty Hall paradox; Crockett gives several other entertaining and embarrassing comments that vos Savant received from so-called experts. Tierney's article tells what Monty Hall himself thought of the fuss—an interesting human-interest angle!

An extensive account of the history of Simpson's paradox is given in Pearl (2009, pp. 174–182), including many attempts by statisticians and philosophers to resolve it without invoking causation. A more recent account, geared for educators, is given in Pearl (2014).

Savage (2009), Julious and Mullee (1994), and Appleton, French, and Vanderpump (1996) give the three real-world examples of Simpson's paradox mentioned in the text (relating to baseball, kidney stones, and smoking, respectively).

Savage's sure-thing principle (Savage, 1954) is treated in Pearl (2016b), and its corrected causal version is derived in Pearl (2009, pp. 181–182).

Versions of Lord's paradox (Lord, 1967) are described in Glymour (2006); Hernández-Díaz, Schisterman, and Hernán (2006); Senn (2006); Wainer (1991); Wainer and Brown (2007). A comprehensive analysis can be found in Pearl (2016a).

Paradoxes invoking counterfactuals are not included in this chapter but are no less intriguing. For a sample, see Pearl (2013).

References

Appleton, D., French, J., and Vanderpump, M. (1996). Ignoring a covariate: An example of Simpson's paradox. *American Statistician* 50: 340–341.

Crockett, Z. (2015). The time everyone "corrected" the world's smartest woman. *Priceonomics*. Available at: http://priceonomics.com/the-time-everyone-corrected-the-worlds-smartest (posted: February 19, 2015).

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In *Methods in Social Epidemiology*. John Wiley and Sons, San Francisco, CA, 393–428.

Grinstead, C. M., and Snell, J. L. (1998). *Introduction to Probability*. 2nd rev. ed. American Mathematical Society, Providence, RI.

Hernández-Díaz, S., Schisterman, E., and Hernán, M. (2006). The birth weight "paradox" uncovered? *American Journal of Epidemiology* 164: 1115–1120.

Julious, S., and Mullee, M. (1994). Confounding and Simpson's paradox. *British Medical Journal* 309: 1480–1481.

Lindley, D. V. (2014). *Understanding Uncertainty*. Rev. ed. John Wiley and Sons, Hoboken, NJ.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin* 68: 304–305.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York, NY.

Pearl, J. (2013). The curse of free-will and paradox of inevitable regret. *Journal of Causal Inference* 1: 255–257.

Pearl, J. (2014). Understanding Simpson's paradox. *American Statistician* 88: 8–13.

Pearl, J. (2016a). Lord's paradox revisited—(Oh Lord! Kumbaya!). *Journal of Causal Inference* 4. doi:10.1515/jci-2016-0021.

Pearl, J. (2016b). The sure-thing principle. *Journal of Causal Inference* 4: 81–86.

Savage, L. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York, NY.

Savage, S. (2009). *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. John Wiley and Sons, Hoboken, NJ.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine* 25: 4334–4344.

Simon, H. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association* 49: 467–479.

Tierney, J. (July 21, 1991). Behind Monty Hall's doors: Puzzle, debate and answer? *New York Times*.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin* 109: 147–151.

Wainer, H., and Brown, L. (2007). Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. Rao C, Sinharay S, editors. *Handbook of Statistics 26: Psychometrics* Vol. 26. North Holland: Elsevier B.V., 893–918.

(move to page 415)

LATE, 255, 395