Book Review

Causal Inference in Statistics: A Primer

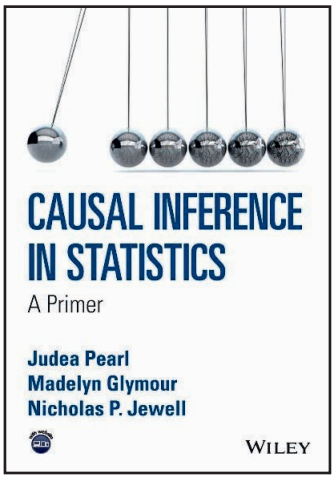
Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell
Wiley, 2016
ISBN: 978-1119186847

Lawrence Stone
Metron, Inc.
Reston, Virginia, USA
stone@metsci.com

If you have ever needed to determine causal effect without performing a randomized controlled trial or if you are tired of hearing statisticians say, "correlation is not causation", then I highly recommend that you read *Causal Inference in Statistics: A Primer* [1] by Judea Pearl, Madelyn Glymour, and Nicholas P. Jewel.

Determining the causal effect of an action is something we do instinctively every day. We know from experience that putting a lit match to a piece of paper will cause the paper to burn. If we do not eat, we will get hungry. When the temperature falls below 32 °F outside, ice may form. We determine causes to guide our actions. For example, determining that malaria is caused by a mosquito rather than "bad air", as its name implies, tells us to use mosquito netting to avoid malaria rather than a gas mask. Knowing that smoking causes lung cancer, we can reduce our risk of getting cancer by not smoking.

Determining causal relationships between actions and results allows us to make intelligent decisions about the actions we should take to avoid risks, improve our health, or safeguard the health of our planet. In many cases, such as the match and the burning paper, the causal relationship is clear. In many others, such as smoking and lung cancer or global warming and extreme weather events, the causal relationship is not so clear.



Judea Pearl is one of the leading developers of the theory of causality, along with Donald B. Rubin [2] and James M. Robins [3]. Pearl is also the recipient of the Association for Computing Machinery's Alan Turing Award for fundamental contributions to probabilistic and causal reasoning. *Causal Inference in Statistics: A Primer* is, in effect, a textbook for a first course in causal inference, complete with study questions (problems) whose answers are available from an instructor's companion website. In level of technical difficulty, this book lies between [4], which is described as a comprehensive exposition of the modern analysis of causation and [5], which is a popular science presentation of Pearl's theory of causal inference. We have used the primer for a study group at Metron (Reston, VA) in which we worked our way through most of the chapters and sections, discussing them and presenting solutions to some of the study problems. Although the presentation in this book is elementary and requires little background, we found it requires a lot of effort to understand the definitions of causality and the methods presented for performing causal inference. Nonetheless, this is an important and developing extension of statistical inference which will become an increasingly significant area of statistical analysis. The well-educated statistician or analyst should at minimum understand the concepts of causal inference and ideally be able to perform causal analysis.

Before reviewing the book, I present some background on causality to introduce the reader to this subject and provide some understanding of the long struggle to develop a satisfactory definition of causality and methods for performing causal inference.

BACKGROUND

Neither classical nor Bayesian statistics provide methods for determining or estimating causality. Statistical methods can estimate correlation, but as we are continually reminded, correlation is not causation. A classic example of this is the data Francis Galton collected on the heights of fathers and sons [6]. He determined that every extra inch in height of the father produced (on average) an extra half-inch of height in the son. He called this relationship the *correlation* between the height of the father and that of his son. This type analysis was later formalized by Karl Pearson into a mathematical method for computing the slope of a (properly rescaled) regression line. Pearson called this slope the correlation coefficient. A peculiar feature of this correlation is that it goes both ways. Tall dads tend to have tall sons, and tall sons tend to have tall dads. The correlation coefficient is the same both ways. Which is the cause and which the effect? Correlation and statistics have no way of telling us. However, we know that it is the father's extra height that tends to produce a taller than average son, not the other way around. We know this because we have in our mind a simple causal model of inheritance which says that the father's height is a cause of the son's height, not the other way around. What Galton's analysis does for us is to quantify the causal relationship. The notions of causality, causal model, and statistical estimates of causal effects have been developed by Pearl and his colleagues into a methodology which allows us to estimate, in certain circumstances, the causal effect of an action on an outcome. Before pressing on to explain Pearl's definition of causality, I briefly review the history of attempts to define causality. Kleinberg and Hripcsak [7] provide an excellent overview of causal inference and various definitions of causality from the perspective of bioinformatics. Researchers in econometrics, bio-

informatics, and epidemiology have been in the forefront of the development and use of causal inference.

DEFINITIONS OF CAUSALITY

REGULARITY DEFINITIONS

In 1739, David Hume proposed the regularity definition of causation which says, in effect, if one type of object (say a flame) always produces a second type of object (say heat), the first object (flame) is the cause of the second (heat). Of course, the difficulty with this definition is that it is really defining a correlation not causation. In 1748, Hume [8] amended his definition to read:

> *We may define a cause to be an object followed by another, and where all the objects similar to the first, are followed by objects similar to the second. Or in other words, where if the first object had not been, the second never had existed.*

This definition depends on the notion of counterfactuals—"where if the first object had not been". That is, it depends on imagining the result of something that did not happen, i.e., a counterfactual. In this definition, Hume assumes that people intuitively understand counterfactual reasoning and that it does not need to be defined. In fact, people often intuitively perform counterfactual reasoning to determine cause and effect in their everyday lives. In subsequent years, most people who considered the question of causality ignored the second sentence in Hume's 1748 definition and concentrated on the first sentence, the regularity part.

INUS CONDITIONS

In many cases there may be multiple factors that produce an effect. Mackie [9] produced an updated version of Hume's regularity definition that allows for multiple causes. He defined a cause as some condition that is perhaps *Insufficient* by itself to produce the effect but is a *Nonredundant* part of a set of conditions that may be *Unnecessary* but are *Sufficient*. These are termed the INUS conditions.

BRADFORD HILL CRITERIA

In 1965, the English statistician Bradford Hill [10] proposed a set of nine criteria to provide epidemiologic evidence of a causal relationship. These criteria were used to demonstrate the connection between cigarette smoking and lung cancer. At one time, these criteria were widely accepted as useful for identifying causal relationships in epidemiological studies. However, a problem with the use of these criteria is that many of them rely on judgment rather than scientific verification.

PROBABILISTIC CAUSALITY

One of the difficulties with the above causality definitions is that they are deterministic. Specifically, they do not allow us to determine quantitatively what fraction of the effect is due to each cause. Probabilistic theories of causality [11], [12], and [13] have been proposed to deal with this problem. The basic idea of these theories is that a cause raises the probability of and occurs before its effect. The condition that a cause C raises the probability of an effect E is defined using conditional probabilities as follows:

$$P(E \mid C) > P(E). \tag{1}$$

The difficulty with this definition is that the conditions of the cause being prior to the effect and the relationship in (1) being true are neither necessary nor sufficient for a causal relationship. A classic example is a falling barometer and rain. The falling barometer occurs before the rain and may be seen as increasing the probability of rain, but it is actually the decreasing air pressure that causes both.

GRANGER CAUSALITY

This definition of causality is usually applied to time series. The approach attempts to find if one variable (coupled with the appropriate time lag) is informative about another. Specifically, let W_t represent the knowledge that is available at time t. Then the time series X at time t is said to be a Granger-cause of the time series Y at some time $t+s$, where $s>0$ if

$$P(Y_{t+s} \mid W_t) \neq P(Y_{t+s} \mid W_t - X_t) \tag{2}$$

where we use $W_t - X_t$ to mean the information contained in W_t with that in X_t removed. The inequality in (2) indicates that X_t contains some information about Y_{t+s} that is not in the rest of the set W_t. Although Granger causality may be useful for predictions, it is not suitable for causality or explanation. As an example, consider that smoking causes both lung cancer and stained fingers, and that the stained fingers usually occur before the cancer. However, we cannot prevent lung cancer by wearing gloves when smoking. The primary type of error that Granger causality produces is to mistake the correlation between common effects of a cause for a causal relationship.

POTENTIAL OUTCOMES AND COUNTERFACTUALS

Let Y represent an outcome such as that of a drug trial and $Y(u)$ represent the outcome for single individual u. Let X be a variable that may affect the outcome of the trial, such as whether the individual u took a specific drug. For example, we could set $X=1$ for taking the drug and $X=0$ for not. The *potential outcome* of the variable $Y(u)$ is the value $Y(u)$ would have taken if $X=x$. This is denoted $Y_{X=x}(u)$. If $X \neq x$ in the trial, then $Y_{X=x}(u)$ is the value $Y(u)$ would have had if (counter to the facts) $X=x$. The crucial assumption here is that such a value exists. Rubin's theory [14] of potential outcomes asserts that this value does exist. The Rubin causal model treats counterfactuals as abstract mathematical objects not derived from a causal model (defined below). In the view of Pearl [5, 280–281], using structural causal models to represent causality relations allows the analyst to clearly visualize and understand their assumptions. By contrast, Pearl claims the purely mathematical assumptions required by Rubin can be difficult to understand and verify.

PEARL'S DEFINITION OF CAUSALITY

In the 1980s, Pearl developed Bayesian Networks and wrote the influential book [15] which decades of analysts have used to model and reason about evidence and uncertainty. Unfortunately, Bayesian networks can estimate only associations, not causality. To estimate causality, Pearl defined three additional notions, causal models, interventions, and counterfactuals.

LADDER OF CAUSATION

In Pearl's ladder of causation [5, 28], he envisions three levels of causal reasoning.

LEVEL 1: ASSOCIATION

On this level, we identify and use regularities in observations. What events are associated with one another? This level allows us to make predictions. For example, what does this poll tell us about the chances of a certain candidate winning an election?

LEVEL 2: INTERVENTION

On this level, one can answer questions such as—*If I give the patient a drug for a certain ailment, how much will that increase his chances of being cured?*

LEVEL 3: COUNTERFACTUALS

On this level, one can answer questions such as—*If I had not taken an aspirin, would my headache have gone away?* Recall that the notion of a counterfactual is crucial to Hume's modified definition of cause: "where if the first object had not been, the second never had existed".

Pearl considers intervention to be a step above association in causal reasoning and counterfactuals a step above that.

We are all familiar with the successes of machine learning algorithms, such as speech recognition and the development of systems such as AlphaGoPlus that can beat the best human Go players. These systems are estimating associations, not causation. For example, speech recognition software recognizes the meaning of ambiguous words by using the meaning associated to that word in the context of the previous words in a sentence. Strategy systems such as AlphaGoPlus are making moves that are associated with positive outcomes in its learning sessions. Thus, machine learning is on the first rung of the causality ladder.

Pearl's insight is that to go up the ladder of causation and estimate the effects of interventions or counterfactuals, we must go beyond classical or Bayesian statistics and beyond systems such as Bayesian networks. To do this, Pearl defines structures called causal models and uses them to give a precise meaning to the terms intervention and counterfactual. A causal model must be added to a standard statistical model to perform causal inference at levels two and three of the ladder of causality. A causal model is an a priori assumption, just like a prior distribution distribution is in Bayesian statistics. Of course, the results of the causal inference will depend on the causal model, just as the results of Bayesian inference will depend on the prior distribution assumed. However, as with Bayesian analysis, the assumptions are explicit for everyone to see, understand, and question if they wish.

CAUSAL MODELS

STRUCTURAL CAUSAL MODEL (SCM)

A structural causal model (SCM) consists of two sets of variables, U and V, and a set F of functions f that assign to each variable in V a value based on the other variables in the model. The variables in U are *exogenous*; they do not depend on any other variables in the model. The variables in V are endogenous and must depend on (be a descendent of) one or more of the exogenous variables. The variables in V can also depend on other endogenous variables.

STRUCTURAL EQUATION MODEL (SEM)

If we know the functions $f \in F$ explicitly, then the SCM becomes a structural equation model (SEM).

GRAPHICAL CAUSAL MODEL (GCM)

Every SCM, M, is associated with a graphical causal model (GCM). The GCM contains one node for each variable in M. If the variable Y in M depends on Z in M, then there is a directed edge from Z to Y. *Causal Inference in Statistics* deals only with SCMs that can be represented by directed acyclic graphs, i.e., directed graphs without loops.

The graphical causal model for ice cream sales given in Figure 1 appears in chapter 3 of [1] and is used to illustrate the concept of intervention and the difference between association and causation. This graph represents an SCM in which

$$U = \{U_X, U_Y, U_Z\}, \quad V = \{X, Y, Z\},$$
$$\text{and } F = \{f_Z(U_Z), f_X(U_X, Z), f_Y(U_Y, Z)\} \tag{3}$$

where f_Z, f_X, f_Y are the functions (possibly unknown) that define Z, X, and Y. For this example, we assume the exogenous variables in U are independent random variables. Note that the temperature Z depends only on the exogenous variable U_Z; X

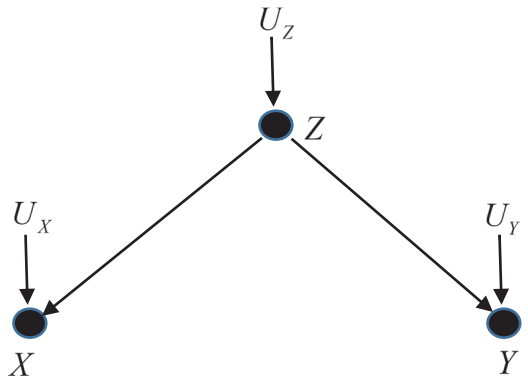


Figure 1
A graphical causal model representing the relationship between temperature Z, ice cream sales X, and crime rates Y. In this graph, X is not independent of Y, but it is conditionally independent of Y given Z.

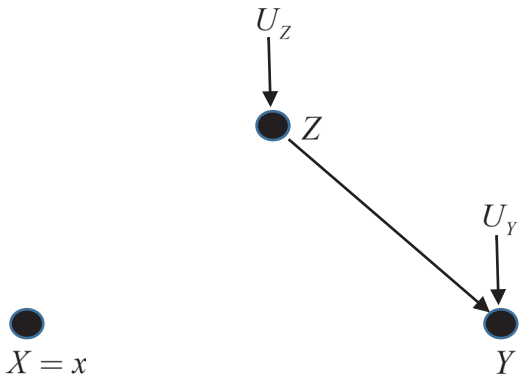


Figure 2
The graphical causal model M_x representing the intervention $X = x$ in the GCM in Figure 1.

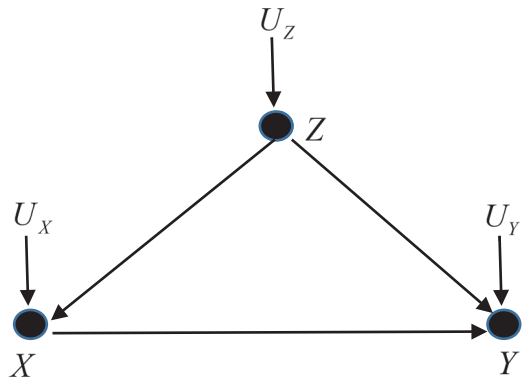


Figure 3
A graphical causal model representing the effects of a new drug, with Z representing gender, X drug usage, and Y recovery.

depends on the endogenous variable Z as well as U_X; and Y depends on Z as well as U_Y. However, there is no dependence of Y, crime rate, on X, ice cream sales. If one performed a statistical analysis of ice cream sales and crime rates, one would likely find a significant correlation between the two. However, from the GCM in Figure 1, we know this is an association not a causal relationship.

If we know the functions f_Z, f_X, f_Y explicitly, then the SCM in (3) becomes an SEM.

INTERVENTIONS AND COUNTERFACTUALS

In order to define interventions and counterfactuals, Pearl first defines the model M_x.

THE MODEL M_x

Suppose we have an SCM M. For defining both an intervention and a counterfactual, we use the SCM model M_x derived from M by setting the endogenous variable X in M to the fixed value x, which we indicate by writing $X = x$.

If M is specified by an SEM, we obtain M_x by replacing X by the fixed value x in all the equations of the model.

If M is specified by a GCM, M_x is obtained by setting the node $X = x$ and removing all the arrows that lead into node X, as shown in Figure 2.

INTERVENTIONS

Pearl uses the notation $P\big(Y \mid do\big(X = x\big)\big)$ to indicate the probability distribution on Y when we intervene to set $X = x$. He points out that this distribution is different than $P(Y \mid X = x)$. He says,

> *In the distributional terminology, $P(Y \mid X = x)$ reflects the population distribution of Y among those individuals whose X value is x. On the other hand, $P\big(Y \mid do\big(X = x\big)\big)$ represents the population distribution of Y if* everyone in the population *had their X value fixed at x.*

To understand what this means, let's consider an SCM in which the variables in U are stochastic and in which we know their joint distribution. We write $P(U = u)$ to denote the probability that the (vector) random variable U has the (vector) value u.

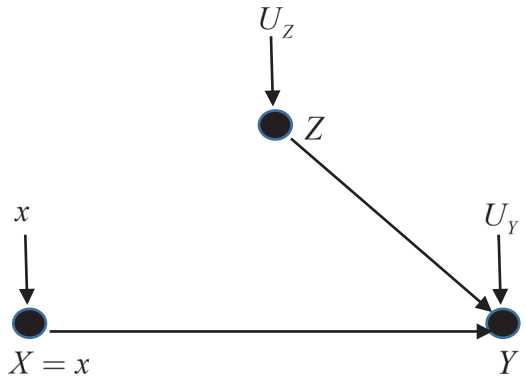


Figure 4
The modified version M_x of the GCM in Figure 3 representing the intervention $X = x$.

To calculate the effect of the intervention $do(X = x)$ on Y, we compute:

$$E\big[Y \mid do\big(X = x\big)\big] \equiv E\big[Y_{X=x}\big] \tag{4}$$

where $Y_{X=x}$ represents Y in the modified model M_x, but the expectation is taken over the unmodified (prior) distribution on U. For simplicity of notation we will often write Y_X for $Y_{X=x}$.

As an example, suppose we have an SEM where the functions f_Z, f_X, f_Y are known explicitly and where:

$$\begin{aligned} Z &= f_Z\big(U_Z\big) = U_Z \\ X &= f_X\big(Z, U_X\big) = g_X\big(Z\big) + U_X \\ Y &= f_Y\big(Z, X, U_Y\big) = g_Y\big(Z, X\big) + U_Y, \end{aligned} \tag{5}$$

i.e., we assume the functions g_X and g_Y are known explicitly. The GCM in Figure 3 shows a graphical version of this model in which Z represents gender, X drug usage, and Y recovery.

Figure 4 shows the graphical model M_x representing the intervention $X = x$.

Using (5), we can calculate the expectation in (4) by

$$E\left[Y \mid do\left(X = x\right)\right] = E\left[Y_x\right] = E\left[g_Y\left(Z, x\right) + U_Y\right] = \int\left[g_Y\left(U_Z, x\right) + U_Y\right]P\left(du\right). \tag{6}$$

In computing the expectation in (6), we have set $X = x$ in the equation for Y and computed the expectation of Y over the prior distribution on U to obtain $E[Y_x]$ and the effect of the intervention $X = x$.

To illustrate the difference between computing a conditional probability and a probability conditioned on an intervention, consider the question of whether a given vaccine tends to protect a person from contracting a disease. If we simply calculate the conditional probability of getting that disease given a person was vaccinated, we are estimating only an association. If the probability is lower for vaccinated than for unvaccinated people, a reason could be that people who tend to get vaccinated are healthier than those who do not or that they tend to have some natural immunity.

A way around this problem is to conduct a randomized controlled trial (RCT). In an RCT, two populations are selected at random. The first receives the vaccine, the other doesn't. The purpose of the randomization is to obtain two populations that are as close to identical as possible so that the only difference between the two is whether people were vaccinated or not. In this case, we can ascribe the decreased chance of contracting the disease to the vaccine. However, in many cases it may be too costly, too difficult, or even unethical to conduct an RCT. For example, you would not want to perform an RCT to determine if smoking causes cancer. For many questions, we are stuck with observational data.

In *Causal Inference in Statistics,* Pearl and his coauthors describe situations and methods by which we can perform causal estimation using observational data. This is the crucial capability provided by Pearl's model of causal inference that is not available from standard statistical techniques.

COUNTERFACTUALS

The counterfactual $P(Y_x = y \mid X = x')$ is the probability of $Y_x = y$ in the model M_x conditioned on the counterfactual $X = x' \neq x$. Recall that Y_x depends on the vector of random variables U. We indicate this by writing $Y_x(u)$ for the value of Y_x when $U=u$.

Let $P(U = u \mid X = x')$ be the probability that $U=u$ given $X = x'$, and let $\mu_{X=x'}$ be the resulting probability distribution on U. Then:

$$P\left(Y_x = y \mid X = x'\right) = E_{\mu_{X=x'}}\left[P\left(Y_x\left(u\right) = y\right)\right] = \int P\left(Y_x\left(u\right) = y\right)P\left(du \mid X = x'\right) \tag{7}$$

where the subscript $\mu_{X=x'}$ on E means that expectation is taken with respect to the measure $\mu_{X=x'}$. More generally:

$$E\left[Y_x \mid X = x'\right] = E_{\mu_{X=x'}}\left[Y_x\right] \tag{8}$$

Furthermore, we can calculate the expectation of Y_x under any counterfactual event $E = e$ by

$$E\left[Y_x \mid E = e\right] = E_{\mu_{E=e}}\left[Y_x\right] \tag{9}$$

COMMENTS ON INTERVENTIONS AND COUNTERFACTUALS

For both interventions and counterfactuals, we compute the expected value of Y_x in model M_x. For interventions, the expectation of Y_x is taken over by the unmodified (prior) distribution on U. For counterfactuals, the expectation of Y_x is taken over by the distribution of U conditioned on the counterfactual, e.g., $X = x' \neq x$.

For counterfactuals, we are estimating the probability of outcomes in an alternate world that does not exist. Surprisingly, there are ways to do this. One way is to have a detailed and accurate model of the system (world) you are analyzing and use that model to compute counterfactual probabilities. Consider the question of whether climate change has increased the probability of extreme weather events. Here we have to consider a counterfactual: if the global average temperature had been 1.5 °F cooler than it is now, how would that have changed the probability of these extreme events occurring? Think about the recent wildfires in Australia at the end of 2019 and beginning of 2020. Is this extreme event now more likely to occur because of climate change?

The special supplement to the Bulletin of the American Meteorological Society of January 2019 [16] includes climate change attribution assessments for seventeen different extreme events from around the world during 2017. For 16 out of those 17 events, the assessments concluded that climate change (global warming) increased the probability of their occurrence. In one of most striking assessments, [17] found that the likelihood of a heat wave at least as hot as the one that happened in the European-Mediterranean region in the summer of 2017 "is at least 3.5 times higher compared to 1950". The probability of this event is now 10% per year.

How was [17] able to make this estimate? The authors developed a statistical model of temperatures in this region. This model depends on the global mean surface temperature (GMST). Using climate models with enough data to analyze the distribution of past and present temperatures, they verified that the GMST influences only the mean of the distribution not the shape. The authors used the temperature distribution resulting from the 2017 value of GMST to determine the probability of experiencing a heat wave in the European-Mediterranean region at least as hot as the one that occurred in 2017. They then compared this to the probability of this event occurring using the (counterfactual) temperature distribution based on the 1950 GMST. The difference or ratio of these two probabilities provides an estimate of the effect of global warming on the probability of occurrence of this extreme event.

As well as providing a quantitative estimate of the effect of global warming on the occurrence of this extreme weather event, this analysis now changes the terms of discussion "from I do not believe in global warming" to "how good is the model that was used to estimate this effect". The latter is a more di-

rected and scientific question and more suitable to quantitative analysis.

Unfortunately, we do not always have a detailed model available to answer our counterfactual questions. Often, we have only observational data. This is where GCMs and Pearl's theory of causality can help us out. If we have a GCM, then in some circumstances (identified in [1]), we can estimate counterfactuals from observational data. Again, the GCM is an assumption, but now the question of the validity of the estimate can be reduced to a more contained and scientific question of whether the GCM is a good model or not. In any case, the assumptions under which the estimate is valid are clearly stated and generally easy to understand.

OUTLINE OF CAUSAL INFLUENCE IN STATISTICS

The book contains four chapters:

1. Preliminaries: Statistical and Causal Models
2. Graphical Models and Their Applications
3. The Effects of Intervention
4. Counterfactuals and Their Applications

CHAPTER 1

Chapter 1 begins with a discussion of why we should study causality and some examples of the fact that standard statistical methods can lead us astray when we use them to estimate causality. The examples are a form of Simpson's paradox, which showed that a statistical association can hold over a whole population but be reversed in every subpopulation. The book presents several examples [1, 3–4] of this paradox, including a hypothetical study of the effects of exercise on cholesterol. If we segregate the data in this study by age, we see that exercise tends to reduce cholesterol levels in every age group; but, if we aggregate the data over all age groups, we find that more exercise tends to produce higher cholesterol levels. The problem here is that older people, who have higher cholesterol levels than younger people, also tend to exercise more. This highlights the question of when to segregate and when to aggregate data when estimating an effect. Statistics by itself has no satisfactory answer to this question. Chapter 1 promises that causal inference will provide an answer to this problem.

The remainder of the chapter covers basic probability, statistics, and graphs. It concludes by defining SCMs and GCMs. (See the section Causal Models.)

CHAPTER 2

Chapter 2 discusses how to use graphs to model dependencies in data. It discusses the notions of chains, forks, and colliders and their importance in causal estimation. It defines the notion of *d-separation* which is an important concept for performing causal estimates. The d-separation property allows one to estimate the effect of an intervention using observational data, i.e., without having to perform an RCT. This chapter also discusses ways in which one can use data to test the validity of a GCM.

CHAPTER 3

Chapter 3 discusses how to estimate the effect of an intervention. Having defined SCMs and GCMs in chapter 2, [1] defines intervention in terms of GCMs as we did in the section Graphical Causal Model (GCM) and defines the "do" operator, e.g., $do(X = x)$ to indicate an intervention. Recall that $P\big(Y \mid do\big(X = x\big)\big)$ is the distribution of the values of Y if every member of the population had their X value set to x. By contrast, $P(Y \mid X = x)$ is the distribution of values of Y among those members of the population whose X value happens to equal x. The latter is an association; the former a causal estimate—what would be the effect on Y of setting $X = x$? The causal estimate is a prediction of the effect produced by the intervention $X = x$. An example is the estimate of the reduction in the probability of contracting a disease if someone is given a vaccination against that disease.

The notion of intervention, although simple to state, is crucial to Pearl's definition of causality. For convenience and emphasis, we repeat it here. First let us recall the definition of an SCM given in the section Structural Causal Model.

An SCM consists of two sets of variables, U and V, and a set F of functions f that assign to each variable in V a value based on the other variables in the model. The variables in U are *exogenous,* they do not depend on any other variables in the model. The variables in V are *endogenous* and must depend on (be a descendent of) one or more of the exogenous variables. The variables in V can also depend on other endogenous variables.

If the equations in F are known explicitly, then the SCM becomes an SEM. Even if we do not know the equations in the SCM explicitly, we can construct a GCM. For example, if we did not have explicit versions of the equations defining the SCM specified by (5), we could represent this SCM by the GCM in Figure 3. The specification of the variables upon which each equation in F depends tells us where to place arrows in the GCM representation of the SCM.

DEFINITION OF CAUSAL EFFECT

Suppose we have a model M specified by an SEM or GCM. (Note any SCM can be represented by a GCM, so we need only consider SEMs and GCMs). Then, M_x is the model obtained by setting the variable $X = x$ in the SEM or the node $X = x$ in the GCM and removing all arrows into X. To calculate the causal effect of the intervention $do(X = x)$ on Y, we compute

$$E\Big[Y \mid do\big(X = x\big)\Big] \equiv E\big[Y_x\big] \tag{10}$$

Let $X = 1$ if a drug is given to a patient and $X = 0$ if not. Suppose Y is the outcome where $Y = 1$ if cured and $Y = 0$ if not. The book defines the average causal effect of the intervention $X = 1$ as

$$E\Big[Y \mid do\big(X = 1\big)\Big] - E\Big[Y \mid do\big(X = 0\big)\Big] \tag{11}$$

In this case, we obtain an estimate of the increased probability of cure when a patient takes the drug compared to not taking the drug. More generally, (11) represents the average causal effect on Y of the intervention $do(X = 1)$ whatever the intervention X represents.

Using the definition of intervention in (10), [1] is able to give guidance on when to segregate or adjust data when performing statistical estimations.

As an example, consider the situation represented by Figure 3. To estimate the effectiveness of the drug, we imagine a hypothetical intervention in which the drug is administered uniformly to everyone in the population and compare the recovery rate to the situation where no one takes the drug. That is, we wish to estimate

$$P\big(Y=1\,|\,do\big(X=1\big)\big)-P\big(Y=1\,|\,do\big(X=0\big)\big) \tag{12}$$

This is the average causal effect defined in (11). However, we cannot simply estimate this effect from observational data because, as we see from Figure 3, our GCM says that gender affects both drug usage and recovery. To estimate the average causal effect, we must change the model M given by Figure 3 to the model M_x given by Figure 4.

Using this model, [1] shows that

$$P\big(Y=y\,|\,do\big(X=x\big)\big)=\sum_z P\big(Y=y\,|\,X=x,Z=z\big)P\big(Z=z\big) \tag{13}$$

Note that the right-hand side of (13) contains only probabilities that can be estimated from observational data. This equation is the adjustment equation which says that in this case, we must adjust our estimates of the effectiveness of the drug conditioned on sex and then produce an overall estimate by weighting these estimates by the distribution of sex in the population under consideration. That is, we are adjusting our estimates by the marginal distribution of Z.

BACKDOOR AND FRONT-DOOR CRITERIA

This chapter defines the notions in GCMs of the backdoor and front-door criteria for a GCM which enable us to use observational data to estimate the effect of an intervention. The backdoor criterion is a special case of d-separation. When a GCM satisfies these criteria, [1] gives formulas for calculating the effects of interventions (e.g., the expectation in (10)) using observational data. This is particularly useful for situations where randomized controlled trials are not feasible or their data are not available. Since the backdoor and front-door criteria apply to GCMs, they give us a way to estimate causal effect when we do not have explicit formulas for the functions in F, i.e., when we do not have an SEM. The chapter describes other methods for estimating the effect of an intervention such as inverse probability weighting and mediation.

CAUSAL INFERENCE IN LINEAR SYSTEMS

The chapter finishes by illustrating causal inference in linear systems and discussing the difference between structural (causal) coefficients in linear systems and regression coefficients. If one has an SCM and performs a regression in accordance with that, the resulting regression coefficients are structural coefficients allowing one to compute causal effects. By contrast, regression coefficients without the context of a SCM give estimates of associations only.

CHAPTER 4

Chapter 4, which deals with counterfactuals, is the most challenging and difficult to absorb. One of the reasons that it was difficult for me is that the chapter does not provide a crisp mathematical definition of counterfactual, such as the one given in the section Interventions and Counterfactuals above.

DIFFERENCE BETWEEN COUNTERFACTUALS AND INTERVENTIONS

The crucial difference between the *do* operator and a counterfactual is that the $do(X = x)$ operator captures the behavior of a population under the intervention $X = x$, whereas $Y_{x'}(u)$ describes the behavior of the individual u under the condition $X = x'$. If it happens that $X = x \neq x'$ in our data for u, then we are estimating a counterfactual, e.g., what would a person's salary have been if she had gone to college rather than beginning work right after high school. We stated this difference in mathematical terms, in the section Interventions and the section Counterfactuals.

COUNTERFACTUAL EXAMPLE

A simple example given in [5, 273–279] illustrates the concept of counterfactuals and how they differ from interventions. Table 8.1 in [5] shows data listing employees, their salary S, education ED (= 0, 1, or 2 for high school, college, or graduate degree), and years of work experience EX. From this table one can perform a linear regression to obtain

$$S = \$65{,}000 + 2{,}500 \times EX + 5000 \times ED \tag{14}$$

as the expected salary of a worker as a function of years of experience and education. In the data, Alice has 6 years of experience, a high school education (ED = 0), and a salary of \$81,000. The counterfactual question is, what would Alice's salary be if she had a college degree, i.e., what is the value of $S_{ED=1}$(Alice)? Using the regression in (14), we could answer this question by setting EX = 6 and ED = 1 to obtain \$85,000. However, the regression in (14) does not account for the fact that education and experience are dependent. We know that

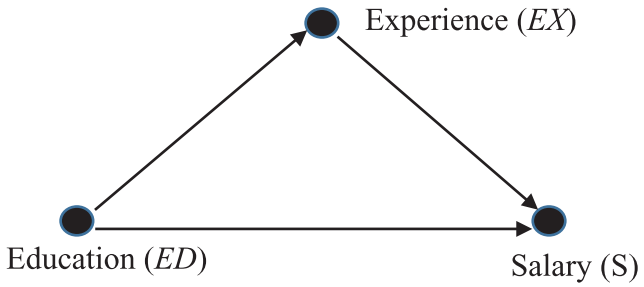


Figure 5
Effect of education (*ED*) and experience (*EX*) on salary *S*.

that the number of years of education tends to lower the number of years of experience. With this in mind, we construct the GCM in Figure 5.

This GCM indicates that education has an effect on experience. It also says that salary does not affect education. We know that causal relationship goes the other way around, namely education affects salary as shown. With this in mind, we rewrite (14) in terms of a SEM, specifically

$$S = \$65{,}000 + \$2{,}500 \times EX + 5000 \times ED + U_S \tag{15}$$

where U_S represent the individual factors that affect a person's salary. Since Figure 5 says that education affects experience, we next perform a linear regression on the data to estimate that effect. By Figure 5, salary does not affect experience, so we set the coefficient of S to 0 in the regression. Suppose we obtain

$$EX = 10 - 4 \times ED + U_{EX} \tag{16}$$

To perform the counterfactual analysis to obtain $S_{ED=1}$(Alice), we first put Alice's experience, education, and salary into (15) and (16) to obtain U_S(Alice) = \$1,000 and U_{EX}(Alice) = −4. We now erase any arrows pointing into ED and set ED=1. In this case there are no arrows pointing into ED, so this step is trivial. In many cases, there are arrows pointing into the counterfactual variable and these must be removed from the model. Finally, we put the values of U_S(Alice) = \$1,000, U_{EX}(Alice) = −4, and ED=1 into the model. First, we compute Alice's experience if she had gone to college. From (16), we see that $EX_{ED=1}$(Alice) = 2. Then from (15) we compute

$$S_{ED=1}(\text{Alice}) = \$65{,}000 + \$2{,}500 \times 2 + \$5{,}000 \times 1 + \$1{,}000 = \$76{,}000$$

as an estimate of Alice's salary if she had gone to college. This is lower than the \$85,000 estimate from the regression. The reason the counterfactual estimate is lower than the regression estimate is that it accounts for the fact that going to college would reduce the number of years of experience that Alice has. In addition, the counterfactual analysis accounts for the terms U_S(Alice) and U_{EX}(Alice) unique to Alice.

As one can see even in this simple deterministic case, counterfactual analysis is not simple.

CONTENTS OF CHAPTER 4

Chapter 4 examines both deterministic and probabilistic counterfactuals. It also provides examples of practical uses of counterfactuals, such as determining the effectiveness of a government program or estimating the effect of sex discrimination in hiring. These examples give us a feeling for the important questions that counterfactual analysis can help us answer. The book finishes with a description of some mathematical tools for estimating probabilities of causation and mediation.

SUMMARY

Modern causal inference, which has developed methods for obtaining quantitative estimates of the effect of interventions and counterfactuals, is an important and relatively new area of analysis. Every analyst should be familiar with the concepts and definitions of causal inference. Causal inference represents a significant extension of standard statistical analysis that should become an increasingly important tool for answering questions about the effectiveness of interventions and for developing artificial intelligence (AI)-like systems that can reason and make decisions. Present AI systems don't reason at the counterfactual level (causality level 3). They make decisions based only on association, unlike humans who can also make decisions based on counterfactual reasoning.

Causal Inference in Statistics is a good introduction to Pearl's version of causal inference. Even though it is a "primer", it requires substantial effort on the reader's part to understand and to apply the concepts and tools presented. The presentation is informal and requires little background beyond basic probability, which is useful for an introduction but frustrating if you are looking for a more mathematical and rigorous approach. For that one has to delve into [4], which can be daunting.

In summary, I highly recommend this book as an introduction to this emerging and important area of analysis. Other introductory texts on causal inference are [2], [3], and [18].

REFERENCES

1. Pearl, J., Glymour, M., and Jewell, N. P. *Causal Inference in Statistics: A Primer*. New York, NY: Wiley, 2016.
2. Imbens, G. W., and Rubin, D. B. *Causal Inference in Statistics, Social and Biomedical Sciences*. Cambridge, UK: Cambridge University Press, 2015.
3. Hernan, M. A., and Robins, J. M. *Causal Inference: What If*. Boca Raton, FL: CRC Press, 2019.
4. Pearl, J. *Causality: Models, Reasoning, and Inference* (2nd ed.). New York, NY: Cambridge University Press, 2009.
5. Pearl, J., and Mackenzie, D. *The Book of Why*. New York, NY: Basic Books, 2018.
6. Galton, F. Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, (1886), 246–263.
7. Kleinberg, S., and Hripcsak, G. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, Vol. 44, 6 (2011), 1102–1112.
8. Hume, D. *An Enquiry Concerning Human Understanding*. 1748.
9. Mackie, J. L. Causes and conditions. *American Philosophical Quarterly*, Vol. 12, (1965), 245–265.
10. Hill, A. B. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, Vol. 58, 5 (1965), 295–300.
11. Good, I. J. A causal calculus (I). *British Journal for the Philosophy of Science*, Vol. XI, 44 (1961), 305–318.
12. Eells, E. *Probabilistic Theory of Causality*. New York, NY: Cambridge University Press, 1991.
13. Suppes, P. A. *Probabilistic Theory of Causality*. New York, NY: North Holland, 1970.

14. Rubin, D. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, Vol. 66, (1974), 688–701.
15. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.
16. Herring, S. C., Christides, N., Hoell, A., Hoerling M., and Stott, P.A. Explaining extreme events of 2017 from a climate perspective. *Bulletin of the American Meteorological Society*, Vol. 100, I (2019), S1–S117.
17. Kew, S. F., Phillip, S. Y., van Oldenborgh, G. J., Otto, F. E. L., Vautard, R., and van der Schrier, G. The exceptional summer heat wave in southern Europe 2017. *Bulletin of the American Meteorological Society*, Vol. 100, 1 (2019), S49–S53.
18. Morgan, S. L. W. C. *Counterfactuals and Causal Inference*. Cambridge, UK: Cambridge University Press, 2015.