# Chapter 1
# A DEFINITION OF CAUSAL EFFECT

By reading this book you are expressing an interest in learning about causal inference. But, as a human being, you have already mastered the fundamental concepts of causal inference. You certainly know what a causal effect is; you clearly understand the difference between association and causation; and you have used this knowledge constantly throughout your life. In fact, had you not understood these causal concepts, you would have not survived long enough to read this chapter—or even to learn to read. As a toddler you would have jumped right into the swimming pool after observing that those who did so were later able to reach the jam jar. As a teenager, you would have skied down the most dangerous slopes after observing that those who did so were more likely to win the next ski race. As a parent, you would have refused to give antibiotics to your sick child after observing that those children who took their medicines were less likely to be playing in the park the next day.

Since you already understand the definition of causal effect and the difference between association and causation, do not expect to gain deep conceptual insights from this chapter. Rather, the purpose of this chapter is to introduce mathematical notation that formalizes the causal intuition that you already possess. Make sure that you can match your causal intuition with the mathematical notation introduced here. This notation is necessary to precisely define causal concepts, and we will use it throughout the book.

## 1.1 Individual causal effects

Zeus is a patient waiting for a heart transplant. On January 1, he receives a new heart. Five days later, he dies. Imagine that we can somehow know, perhaps by divine revelation, that had Zeus not received a heart transplant on January 1, he would have been alive five days later. Equipped with this information most would agree that the transplant caused Zeus's death. The heart transplant intervention had a causal effect on Zeus's five-day survival.

Another patient, Hera, also received a heart transplant on January 1. Five days later she was alive. Imagine we can somehow know that, had Hera not received the heart on January 1, she would still have been alive five days later. Hence the transplant did not have a causal effect on Hera's five-day survival.

These two vignettes illustrate how humans reason about causal effects: We compare (usually only mentally) the outcome when an action $A$ is taken with the outcome when the action $A$ is withheld. If the two outcomes differ, we say that the action $A$ has a causal effect, causative or preventive, on the outcome. Otherwise, we say that the action $A$ has no causal effect on the outcome. Epidemiologists, statisticians, economists, and other social scientists often refer to the action $A$ as an intervention, an exposure, or a treatment.

Capital letters represent random variables. We assume subjects are independent and identically distributed and thus suppress the individual subscript $i$ in $A_i$ and the other variables.
Lower case letters denote particular values of a random variable.

To make our causal intuition amenable to mathematical and statistical analysis we will introduce some notation. Consider a dichotomous treatment variable $A$ (1: treated, 0: untreated) and a dichotomous outcome variable $Y$ (1: death, 0: survival). In this book we refer to variables such as $A$ and $Y$ that may have different values for different individuals or subjects as *random variables*. Let $Y^{a=1}$ (read $Y$ under treatment $a = 1$) be the outcome variable that would have been observed under the treatment value $a = 1$, and $Y^{a=0}$ (read $Y$ under treatment $a = 0$) the outcome variable that would have been

We abbreviate the expression "individual $i$ has outcome $Y = 1$" by writing $Y_i = 1$ , and analogously for other random variables.

observed under the treatment value $a = 0$. $Y^{a=1}$ and $Y^{a=0}$ are also random variables. Zeus has $Y^{a=1} = 1$ and $Y^{a=0} = 0$ because he died when treated but would have survived if untreated, while Hera has $Y^{a=1} = 0$ and $Y^{a=0} = 0$ because she survived when treated and would also have survived if untreated.

We can now provide a formal definition of a *causal effect for an individual*: the treatment $A$ has a causal effect on an individual's outcome $Y$ if $Y^{a=1} \neq Y^{a=0}$ for the individual. Thus the treatment has a causal effect on Zeus's outcome because $Y^{a=1} = 1 \neq 0 = Y^{a=0}$, but not on Hera's outcome because $Y^{a=1} = 0 = Y^{a=0}$. The variables $Y^{a=1}$ and $Y^{a=0}$ are referred to as *potential outcomes* or as *counterfactual outcomes*. Some authors prefer the term "potential outcomes" to emphasize that, depending on the treatment that is received, either of these two outcomes can be potentially observed. Other authors prefer the term "counterfactual outcomes" to emphasize that these outcomes represent situations that may not actually occur (that is, counter to the fact situations).

Causal effect for individual $i$:
$Y_i^{a=1} \neq Y_i^{a=0}$

For each subject, one of the counterfactual outcomes—the one that corresponds to the treatment value that the subject actually received—is actually factual. For example, because Zeus was actually treated ($A = 1$), his counterfactual outcome under treatment $Y^{a=1} = 1$ is equal to his observed (actual) outcome $Y = 1$. That is, a subject with observed treatment $A$ equal to $a$, has observed outcome $Y$ equal to his counterfactual outcome $Y^a$. This equality can be succinctly expressed as $Y = Y^A$ where $Y^A$ denotes the counterfactual $Y^a$ evaluated at the value $a$ corresponding to the subject's observed treatment $A$. The equality $Y = Y^A$ is referred to as *consistency*.

Consistency:
if $A_i = a$, then $Y_i^a = Y^{A_i} = Y_i$

Individual causal effects are defined as a contrast of the values of counterfactual outcomes, but only one of those outcomes is observed for each individual—the one corresponding to the treatment value actually experienced by the subject. All other counterfactual outcomes remain unobserved. The unhappy conclusion is that, in general, individual causal effects cannot be identified, i.e., computed from the observed data, because of missing data. (See Fine Point 2.1 for a possible exception.)

## 1.2 Average causal effects

We needed three pieces of information to define an individual causal effect: an outcome of interest, the actions $a = 1$ and $a = 0$ to be compared, and the individual whose counterfactual outcomes $Y^{a=0}$ and $Y^{a=1}$ are to be compared. However, because identifying individual causal effects is generally not possible, we now turn our attention to an aggregated causal effect: the average causal effect in a population of individuals. To define it, we need three pieces of information: an outcome of interest, the actions $a = 1$ and $a = 0$ to be compared, and a well defined population of individuals whose outcomes $Y^{a=0}$ and $Y^{a=1}$ are to be compared.

Take Zeus's extended family as our population of interest. Table 1.1 shows the counterfactual outcomes under both treatment ($a = 1$) and no treatment ($a = 0$) for all 20 members of our population. Let us first focus our attention on the last column: the outcome $Y^{a=1}$ that would have been observed for each individual if they had received the treatment (a heart transplant). Half of the members of the population (10 out of 20) would have died if they had received a heart transplant. That is, the proportion of individuals that would have developed the outcome had all population subjects received treatment

Fine Point 1.1

**Interference between subjects.** An implicit assumption in our definition of counterfactual outcome is that a subject's counterfactual outcome under treatment value $a$ does not depend on other subjects' treatment values. For example, we implicitly assumed that Zeus would die if he received a heart transplant, regardless of whether Hera also received a heart transplant. That is, Hera's treatment value did not interfere with Zeus's outcome. On the other hand, suppose that Hera's getting a new heart upsets Zeus to the extent that he would not survive his own heart transplant, even though he would have survived had Hera not been transplanted. In this scenario, Hera's treatment interferes with Zeus's outcome. Interference between subjects is common in studies that deal with contagious agents or educational programs, in which an individual's outcome is influenced by their social interaction with other population members. In the presence of interference, the counterfactual $Y_i^a$ for an individual $i$ is not well defined because an individual's outcome depends also on other individuals' treatment values. As a consequence "the causal effect of heart transplant on Zeus's outcome" is not well defined when there is interference. Rather, one needs to refer to "the causal effect of heart transplant on Zeus's outcome when Hera does not get a new heart" or "the causal effect of heart transplant on Zeus's outcome when Hera does get a new heart." If other relatives and friends' treatment also interfere with Zeus's outcome, then one may need to refer to the causal effect of heart transplant on Zeus's outcome when "no relative or friend gets a new heart," "when only Hera gets a new heart," etc. because the causal effect of treatment on Zeus's outcome may differ for each particular allocation of hearts. The assumption of no interference was labeled "no interaction between units" by Cox (1958), and is included in the "stable-unit-treatment-value assumption (SUTVA)" described by Rubin (1980). Unless otherwise specified, we will assume no interference throughout this book.

Table 1.1

| | $Y^{a=0}$ | $Y^{a=1}$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 0 | 0 |
| Poseidon | 1 | 0 |
| Hera | 0 | 0 |
| Zeus | 0 | 1 |
| Artemis | 1 | 1 |
| Apollo | 1 | 0 |
| Leto | 0 | 1 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 0 | 1 |
| Aphrodite | 0 | 1 |
| Cyclope | 0 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

Average causal effect in population: $\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$

$a = 1$ is $\Pr[Y^{a=1} = 1] = 10/20 = 0.5$. Similarly, from the other column of Table 1.1, we can conclude that half of the members of the population (10 out of 20) would have died if they had not received a heart transplant. That is, the proportion of subjects that would have developed the outcome had all population subjects received no treatment $a = 0$ is $\Pr[Y^{a=0} = 1] = 10/20 = 0.5$. Note that we have computed the counterfactual risk under treatment to be 0.5 by counting the number of deaths (10) and dividing them by the total number of individuals (20), which is the same as computing the average of the counterfactual outcome across all individuals in the population (if you do not see the equivalence between risk and average for a dichotomous outcome, please use the data in Table 1.1 to compute the average of $Y^{a=1}$).

We are now ready to provide a formal definition of the *average causal effect* in the population: an average causal effect of treatment $A$ on outcome $Y$ is present if $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ in the population of interest. Under this definition, treatment $A$ does not have an average causal effect on outcome $Y$ in our population because both the risk of death under treatment $\Pr[Y^{a=1} = 1]$ and the risk of death under no treatment $\Pr[Y^{a=0} = 1]$ are 0.5. That is, it does not matter whether all or none of the individuals receive a heart transplant: half of them would die in either case. When, like here, the average causal effect in the population is null, we say that the *null hypothesis of no average causal effect* is true. Because the risk equals the average and because the letter E is usually employed to represent the population average or mean (also referred to as 'E'xpectation), we can rewrite the definition of a non-null average causal effect in the population as $\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$ so that the definition applies to both dichotomous and nondichotomous outcomes.

The presence of an "average causal effect of heart transplant $A$" is defined by a contrast that involves the two actions "receiving a heart transplant ($a = 1$)" and "not receiving a heart transplant ($a = 0$)." When more than two actions are possible (i.e., the treatment is not dichotomous), the particular

Fine Point 1.2

**Multiple versions of treatment.** Another implicit assumption in our definition of a subject's counterfactual outcome under treatment value $a$ is that there is only one version of treatment value $A = a$. For example, we said that Zeus would die if he received a heart transplant. This statement implicitly assumes that all heart transplants are performed by the same surgeon using the same procedure and equipment. That is, that there is only one version of the treatment "heart transplant." If there were multiple versions of treatment (e.g., surgeons with different skills), then it is possible that Zeus would survive if his transplant were performed by Asclepios, and would die if his transplant were performed by Hygieia. In the presence of multiple versions of treatment, the counterfactual $Y_i^a$ for an individual $i$ is not well defined because an individual's outcome depends on the version of treatment $a$. As a consequence "the causal effect of heart transplant on Zeus's outcome" is not well defined when there are multiple versions of treatment. Rather, one needs to refer to "the causal effect of heart transplant on Zeus's outcome when Asclepios performs the surgery" or "the causal effect of heart transplant on Zeus's outcome when Hygieia performs the surgery." If other components of treatment (e.g., procedure, place) are also relevant to the outcome, then one may need to refer to "the causal effect of heart transplant on Zeus's outcome when Asclepios performs the surgery using his rod at the temple of Kos" because the causal effect of treatment on Zeus's outcome may differ for each particular version of treatment. The assumption of no multiple versions of treatment is included in the "stable-unit-treatment-value assumption (SUTVA)" described by Rubin (1980). VanderWeele (2009) formalized the weaker assumption of "treatment variation irrelevance," i.e., the assumption that multiple versions of treatment $A = a$ may exist but they all result in the same outcome $Y_i^a$. Unless otherwise specified, we will assume treatment variation irrelevance throughout this book. See Chapter 3 for an extended discussion of this issue.

contrast of interest needs to be specified. For example, "the causal effect of aspirin" is meaningless unless we specify that the contrast of interest is, say, "taking, while alive, 150 mg of aspirin by mouth (or nasogastric tube if need be) daily for 5 years" versus "not taking aspirin." Note that this causal effect is well defined even if counterfactual outcomes under other interventions are not well defined or even do not exist (e.g., "taking, while alive, 500 mg of aspirin by absorption through the skin daily for 5 years").

Absence of an average causal effect does not imply absence of individual effects. In fact, Table 1.1 shows that treatment has an individual causal effect on the outcomes of 12 members (including Zeus) of the population because, for each of these 12 individuals, the value of their counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$ differ. Six of the twelve (including Zeus) were harmed by treatment $\left(Y^{a=1} - Y^{a=0} = 1\right)$; an equal number were helped $\left(Y^{a=1} - Y^{a=0} = -1\right)$. This equality is not an accident: the average causal effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ is always equal to the average $\mathrm{E}[Y^{a=1} - Y^{a=0}]$ of the individual causal effects $Y^{a=1} - Y^{a=0}$, as a difference of averages is equal to the average of the differences. When there is no causal effect for any individual in the population, i.e., $Y^{a=1} = Y^{a=0}$ for all subjects, we say that the *sharp causal null hypothesis* is true. The sharp causal null hypothesis implies the null hypothesis of no average effect.

As discussed in the next chapters, average causal effects *can* sometimes be identified from data, even if individual causal effects cannot. Hereafter we refer to 'average causal effects' simply as 'causal effects' and the null hypothesis of no average effect as the causal null hypothesis. We next describe different measures of the magnitude of a causal effect.

Technical Point 1.1

**Causal effects in the population.** Let $\mathrm{E}[Y^a]$ be the mean counterfactual outcome had all subjects in the population received treatment level $a$. For discrete outcomes, the mean or expected value $\mathrm{E}[Y^a]$ is defined as the weighted sum $\sum_y y\, p_{Y^a}(y)$ over all possible values $y$ of the random variable $Y^a$, where $p_{Y^a}(\cdot)$ is the probability mass function of $Y^a$, i.e., $p_{Y^a}(y) = \Pr[Y^a = y]$. For dichotomous outcomes, $\mathrm{E}[Y^a] = \Pr[Y^a = 1]$. For continuous outcomes, the expected value $\mathrm{E}[Y^a]$ is defined as the integral $\int y f_{Y^a}(y)\, dy$ over all possible values $y$ of the random variable $Y^a$, where $f_{Y^a}(\cdot)$ is the probability density function of $Y^a$. A common representation of the expected value that applies to both discrete and continuous outcomes is $\mathrm{E}[Y^a] = \int y dF_{Y^a}(y)$, where $F_{Y^a}(\cdot)$ is the cumulative distribution function (CDF) of the random variable $Y^a$. We say that there is a non-null average causal effect in the population if $\mathrm{E}[Y^a] \neq \mathrm{E}[Y^{a'}]$ for any two values $a$ and $a'$.

The average causal effect, defined by a contrast of means of counterfactual outcomes, is the most commonly used population causal effect. However, a population causal effect may also be defined as a contrast of, say, medians, variances, hazards, or CDFs of counterfactual outcomes. In general, a causal effect can be defined as a contrast of any functional of the distributions of counterfactual outcomes under different actions or treatment values. The causal null hypothesis refers to the particular contrast of functionals (mean, median, variance, hazard, CDF, ...) used to define the causal effect.

## 1.3 Measures of causal effect

We have seen that the treatment 'heart transplant' $A$ does not have a causal effect on the outcome 'death' $Y$ in our population of 20 family members of Zeus. The causal null hypothesis holds because the two counterfactual risks $\Pr[Y^{a=1} = 1]$ and $\Pr[Y^{a=0} = 1]$ are equal to 0.5. There are equivalent ways of representing the causal null. For example, we could say that the risk $\Pr[Y^{a=1} = 1]$ minus the risk $\Pr\left[Y^{a=0} = 1\right]$ is zero $(0.5 - 0.5 = 0)$ or that the risk $\Pr[Y^{a=1} = 1]$ divided by the risk $\Pr\left[Y^{a=0} = 1\right]$ is one $(0.5/0.5 = 1)$. That is, we can represent the causal null by

**(i)** $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] = 0$

**(ii)** $\dfrac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]} = 1$

**(iii)** $\dfrac{\Pr[Y^{a=1} = 1]/\Pr[Y^{a=1} = 0]}{\Pr[Y^{a=0} = 1]/\Pr[Y^{a=0} = 0]} = 1$

where the left-hand side of the equalities (i), (ii), and (iii) is the causal risk difference, risk ratio, and odds ratio, respectively.

Suppose now that another treatment $A$, cigarette smoking, has a causal effect on another outcome $Y$, lung cancer, in our population. The causal null hypothesis does not hold: $\Pr[Y^{a=1} = 1]$ and $\Pr[Y^{a=0} = 1]$ are not equal. In this setting, the causal risk difference, risk ratio, and odds ratio are not 0, 1, and 1, respectively. Rather, these causal parameters quantify the strength of the same causal effect on different scales. Because the causal risk difference, risk ratio, and odds ratio (and other summaries) measure the causal effect, we refer to them as *effect measures*.

Each effect measure may be used for different purposes. For example, imagine a large population in which 3 in a million individuals would develop the outcome if treated, and 1 in a million individuals would develop the outcome if untreated. The causal risk ratio is 3, and the causal risk difference is 0.000002. The causal risk ratio (multiplicative scale) is used to compute how many times

Fine Point 1.3

**Number needed to treat.** Consider a population of 100 million patients in which 20 million would die within five years if treated ($a = 1$), and 30 million would die within five years if untreated ($a = 0$). This information can be summarized in several equivalent ways:

- the causal risk difference is $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] = 0.2 - 0.3 = -0.1$
- if one treats the 100 million patients, there will be 10 million fewer deaths than if one does not treat those 100 million patients.
- one needs to treat 100 million patients to save 10 million lives
- on average, one needs to treat 10 patients to save 1 life

We refer to the average number of individuals that need to receive treatment $a = 1$ to reduce the number of cases $Y = 1$ by one as the number needed to treat (NNT). In our example the NNT is equal to 10. For treatments that reduce the average number of cases (i.e., the causal risk difference is negative), the NNT is equal to the reciprocal of the absolute value of the causal risk difference:

$$NNT = \frac{-1}{\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]}$$

Like the causal risk difference, the NNT applies to the population and time interval on which it is based. For treatments that increase the average number of cases (i.e., the causal risk difference is positive), one can symmetrically define the *number needed to harm*. The NNT was introduced by Laupacis, Sackett, and Roberts (1988). For a discussion of the relative advantages and disadvantages of the NNT as an effect measure, see Grieve (2003).

treatment, relative to no treatment, increases the disease risk. The causal risk difference (additive scale) is used to compute the absolute number of cases of the disease attributable to the treatment. The use of either the multiplicative or additive scale will depend on the goal of the inference.

## 1.4 Random variability

At this point you could complain that our procedure to compute effect measures is somewhat implausible. Not only did we ignore the well known fact that the immortal Zeus cannot die, but—more to the point—our population in Table 1.1 had only 20 individuals. The populations of interest are typically much larger.

In our tiny population, we collected information from all the subjects. In practice, investigators only collect information on a sample of the population of interest. Even if the counterfactual outcomes of all study subjects were known, working with samples prevents one from obtaining the exact proportion of subjects in the population who had the outcome under treatment value $a$, e.g., the probability of death under no treatment $\Pr[Y^{a=0} = 1]$ cannot be directly computed. One can only estimate this probability.

1st source of random error: Sampling variability

Consider the subjects in Table 1.1. We have previously viewed them as forming a twenty-subject population. Suppose we view them as a random sample from a much larger, near-infinite super-population (e.g., all immortals). We denote the proportion of subjects in the sample who would have died if unex-

posed as $\widehat{\Pr}[Y^{a=0}=1] = 10/20 = 0.50$. The sample proportion $\widehat{\Pr}[Y^{a=0}=1]$ does not have to be exactly equal to the proportion of subjects who would have died if the entire super-population had been unexposed, $\Pr[Y^{a=0}=1]$. For example, suppose $\Pr[Y^{a=0}=1] = 0.57$ in the population but, because of random error due to sampling variability, $\widehat{\Pr}[Y^{a=0}=1] = 0.5$ in our particular sample. We use the sample proportion $\widehat{\Pr}[Y^a=1]$ to estimate the super-population probability $\Pr[Y^a=1]$ under treatment value $a$. The "hat" over Pr indicates that the sample proportion $\widehat{\Pr}[Y^a=1]$ is an estimator of the corresponding population quantity $\Pr[Y^a=1]$. We say that $\widehat{\Pr}[Y^a=1]$ is a *consistent estimator* of $\Pr[Y^a=1]$ because the larger the number of subjects in the sample, the smaller the difference between $\widehat{\Pr}[Y^a=1]$ and $\Pr[Y^a=1]$ is expected to be. This occurs because the error due to sampling variability is random and thus obeys the law of large numbers.

An estimator $\hat{\theta}$ of $\theta$ is consistent if, with probability approaching 1, the difference $\hat{\theta}-\theta$ approaches zero as the sample size increases towards infinity.

Caution: the term 'consistency' when applied to estimators has a different meaning from that which it has when applied to counterfactual outcomes.

Because the super-population probabilities $\Pr[Y^a=1]$ cannot be computed, only consistently estimated by the sample proportions $\widehat{\Pr}[Y^a=1]$, one cannot conclude with certainty that there is, or there is not, a causal effect. Rather, a statistical procedure must be used to test the causal null hypothesis $\Pr[Y^{a=1}=1] = \Pr[Y^{a=0}=1]$; the procedure quantifies the chance that the difference $\widehat{\Pr}[Y^{a=1}=1]$ and $\widehat{\Pr}[Y^{a=0}=1]$ is wholly due to sampling variability.

$2^{\text{nd}}$ source of random error: Nondeterministic counterfactuals

So far we have only considered sampling variability as a source of random error. But there may be another source of random variability: perhaps the values of an individual's counterfactual outcomes are not fixed in advance. We have defined the counterfactual outcome $Y^a$ as the subject's outcome had he received treatment value $a$. For example, in our first vignette, Zeus would have died if treated and would have survived if untreated. As defined, the values of the counterfactual outcomes are fixed or deterministic for each subject, e.g., $Y^{a=1}=1$ and $Y^{a=0}=0$ for Zeus. In other words, Zeus has a 100% chance of dying if treated and a 0% chance of dying if untreated. However, we could imagine another scenario in which Zeus has a 90% chance of dying if treated, and a 10% chance of dying if untreated. In this scenario, the counterfactual outcomes are stochastic or nondeterministic because Zeus's probabilities of dying under treatment (0.9) and under no treatment (0.1) are neither zero or one. The values of $Y^{a=1}$ and $Y^{a=0}$ shown in Table 1.1 would be possible realizations of "random flips of mortality coins" with these probabilities. Further, one would expect that these probabilities vary across subjects because not all subjects are equally susceptible to develop the outcome. Quantum mechanics, in contrast to classical mechanics, holds that outcomes are inherently nondeterministic. That is, if the quantum mechanical probability of Zeus dying is 90%, the theory holds that no matter how much data we collect about Zeus, the uncertainty about whether Zeus will actually develop the outcome if treated is irreducible and statistical methods are needed to quantify it.

Table 1.2

| | $A$ | $Y$ |
|---|---|---|
| Rheia | 0 | 0 |
| Kronos | 0 | 1 |
| Demeter | 0 | 0 |
| Hades | 0 | 0 |
| Hestia | 1 | 0 |
| Poseidon | 1 | 0 |
| Hera | 1 | 0 |
| Zeus | 1 | 1 |
| Artemis | 0 | 1 |
| Apollo | 0 | 1 |
| Leto | 0 | 0 |
| Ares | 1 | 1 |
| Athena | 1 | 1 |
| Hephaestus | 1 | 1 |
| Aphrodite | 1 | 1 |
| Cyclope | 1 | 1 |
| Persephone | 1 | 1 |
| Hermes | 1 | 0 |
| Hebe | 1 | 0 |
| Dionysus | 1 | 0 |

Thus statistics is necessary in causal inference to quantify random error from sampling variability, nondeterministic counterfactuals, or both. However, for pedagogic reasons, we will continue to largely ignore statistical issues until Chapter 10. Specifically, we will assume that counterfactual outcomes are deterministic and that we have recorded data on every subject in a very large (perhaps hypothetical) super-population. This is equivalent to viewing our population of 20 subjects as a population of 20 billion subjects in which 1 billion subjects are identical to Zeus, 1 billion subjects are identical to Hera, and so on. Hence, until Chapter 10, we will carry out our computations with Olympian certainty.

Technical Point 1.2

**Nondeterministic counterfactuals.** For nondeterministic counterfactual outcomes, the mean outcome under treatment value $a$, $\mathrm{E}[Y^a]$, equals the weighted sum $\sum_y y\, p_{Y^a}(y)$ over all possible values $y$ of the random variable $Y^a$, where the probability mass function $p_{Y^a}(\cdot) = \mathrm{E}\left[Q_{Y^a}(\cdot)\right]$, and $Q_{Y^a}(y)$ is a random probability of having outcome $Y = y$ under treatment level $a$. In the example described in the text, $Q_{Y^{a=1}}(1) = 0.9$ for Zeus. (For continuous outcomes, the weighted sum is replaced by an integral.)

More generally, a nondeterministic definition of counterfactual outcome does not attach some particular value of the random variable $Y^a$ to each subject, but rather a statistical distribution $\Theta_{Y^a}(\cdot)$ of $Y^a$. The nondeterministic definition of causal effect is a generalization of the deterministic definition in which $\Theta_{Y^a}(\cdot)$ is a random CDF that may take values between 0 and 1. The average counterfactual outcome in the population $\mathrm{E}[Y^a]$ equals $\mathrm{E}\left\{\mathrm{E}\left[Y^a \mid \Theta_{Y^a}(\cdot)\right]\right\}$. Therefore, $\mathrm{E}[Y^a] = \mathrm{E}\left[\int y\, d\Theta_{Y^a}(y)\right] = \int y\, d\mathrm{E}[\Theta_{Y^a}(y)] = \int y\, dF_{Y^a}(y)$, because we define $F_{Y^a}(\cdot) = \mathrm{E}\left[\Theta_{Y_i^a}(\cdot)\right]$. Although the possibility of nondeterministic counterfactual outcomes implies no changes in our definitions of population causal effect and of effect measures, nondeterministic counterfactual outcomes introduce random variability. This additional variability has implications for the computation of confidence intervals for the effect measures (Robins 1988), as discussed in Chapter 10.

## 1.5 Causation versus association

Obviously, the data available from actual studies look different from those shown in Table 1.1. For example, we would not usually expect to learn Zeus's outcome if treated $Y^{a=1}$ *and* also Zeus's outcome if untreated $Y^{a=0}$. In the real world, we only get to observe one of those outcomes because Zeus is either treated or untreated. We referred to the observed outcome as $Y$. Thus, for each individual, we know the observed treatment level $A$ and the outcome $Y$ as in Table 1.2.

The data in Table 1.2 can be used to compute the proportion of subjects that developed the outcome $Y$ among those subjects in the population that happened to receive treatment value $a$. For example, in Table 1.2, 7 subjects died ($Y = 1$) among the 13 individuals that were treated ($A = 1$). Thus the risk of death in the treated, $\Pr[Y = 1|A = 1]$, was 7/13. In general, we define the conditional probability $\Pr[Y = 1|A = a]$ as the proportion of subjects that developed the outcome $Y$ among those subjects in the population of interest that happened to receive treatment value $a$.

When the proportion of subjects who develop the outcome in the treated $\Pr[Y = 1|A = 1]$ equals the proportion of subjects who develop the outcome in the untreated $\Pr[Y = 1|A = 0]$, we say that treatment $A$ and outcome $Y$ are independent, that $A$ is not associated with $Y$, or that $A$ does not predict $Y$. *Independence* is represented by $Y \coprod A$—or, equivalently, $A \coprod Y$— which is read as $Y$ and $A$ are independent. Some equivalent definitions of independence are

Dawid (1979) introduced the symbol Ш to denote independence

**(i)** $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0] = 0$

**(ii)** $\dfrac{\Pr[Y = 1|A = 1]}{\Pr[Y = 1|A = 0]} = 1$

**(iii)** $\dfrac{\Pr[Y = 1|A = 1]/\Pr[Y = 0|A = 1]}{\Pr[Y = 1|A = 0]/\Pr[Y = 0|A = 0]} = 1$

where the left-hand side of the inequalities (i), (ii), and (iii) is the associational risk difference, risk ratio, and odds ratio, respectively.

For a continuous outcome $Y$ we define *mean independence* between treatment and outcome as: $\mathrm{E}[Y|A=1] = \mathrm{E}[Y|A=0]$. Independence and mean independence are the same concept for dichotomous outcomes.

We say that treatment $A$ and outcome $Y$ are dependent or associated when $\Pr[Y=1|A=1] \neq \Pr[Y=1|A=0]$. In our population, treatment and outcome are indeed associated because $\Pr[Y=1|A=1] = 7/13$ and $\Pr[Y=1|A=0] = 3/7$. The associational risk difference, risk ratio, and odds ratio (and other measures) quantify the strength of the association when it exists. They measure the association on different scales, and we refer to them as *association measures*. These measures are also affected by random variability. However, until Chapter 10, we will disregard statistical issues by assuming that the population in Table 1.2 is extremely large.

For dichotomous outcomes, the risk equals the average in the population, and we can therefore rewrite the definition of association in the population as $\mathrm{E}[Y|A=1] \neq \mathrm{E}[Y|A=0]$. For continuous outcomes $Y$, we can also define association as $\mathrm{E}[Y|A=1] \neq \mathrm{E}[Y|A=0]$. Under this definition, association is essentially the same as the statistical concept of *correlation* between $A$ and a continuous $Y$.

In our population of 20 individuals, we found $(i)$ no causal effect after comparing the risk of death if all 20 individuals had been treated with the risk of death if all 20 individuals had been untreated, and $(ii)$ an association after comparing the risk of death in the 13 individuals who happened to be treated with the risk of death in the 7 individuals who happened to be untreated. Figure 1.1 depicts the causation-association difference. The population (represented by a diamond) is divided into a white area (the treated) and a smaller grey area (the untreated). The definition of causation implies a contrast between the whole white diamond (all subjects treated) and the whole grey diamond (all subjects untreated), whereas association implies a contrast between the white (the treated) and the grey (the untreated) areas of the original diamond.

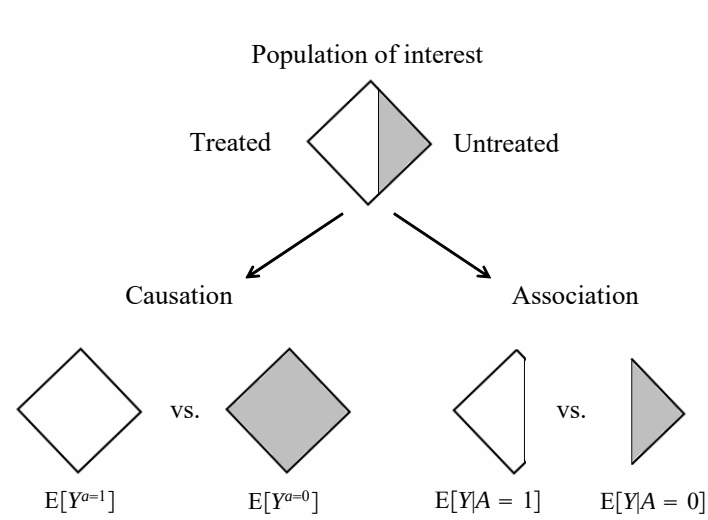


Figure 1.1

We can use the notation we have developed thus far to formalize the distinction between causation and association. The risk $\Pr[Y=1|A=a]$ is a conditional probability: the risk of $Y$ in the subset of the population that meet the condition 'having actually received treatment value $a$' (i.e., $A=a$). In contrast the risk $\Pr[Y^a=1]$ is an unconditional—also known as marginal—probability, the risk of $Y^a$ in the entire population. Therefore, *association* is defined by a different risk in two disjoint subsets of the population determined

The difference between association and causation is critical. Suppose the causal risk ratio of 5-year mortality is 0.5 for aspirin vs. no aspirin, and the corresponding associational risk ratio is 1.5. After a physician learns these results, she decides to withhold aspirin from her patients because those treated with aspirin have a greater risk of dying compared with the untreated. The doctor will be sued for malpractice.

by the subjects' actual treatment value ($A = 1$ or $A = 0$), whereas *causation* is defined by a different risk in the entire population under two different treatment values ($a = 1$ or $a = 0$). Throughout this book we often use the redundant expression 'causal effect' to avoid confusions with a common use of 'effect' meaning simply association.

These radically different definitions explain the well-known adage "association is not causation." In our population, there was association because the mortality risk in the treated (7/13) was greater than that in the untreated (3/7). However, there was no causation because the risk if everybody had been treated (10/20) was the same as the risk if everybody had been untreated. This discrepancy between causation and association would not be surprising if those who received heart transplants were, on average, sicker than those who did not receive a transplant. In Chapter 7 we refer to this discrepancy as *confounding*.

Causal inference requires data like the hypothetical data in Table 1.1, but all we can ever expect to have is real world data like those in Table 1.2. The question is then under which conditions real world data can be used for causal inference. The next chapter provides one answer: conduct a randomized experiment.