Chapter 2
RANDOMIZED EXPERIMENTS

Does your looking up at the sky make other pedestrians look up too? This question has the main components of any causal question: we want to know whether an action (your looking up) affects an outcome (other people's looking up) in a specific population (say, residents of Madrid in 2019). Suppose we challenge you to design a scientific study to answer this question. "Not much of a challenge," you say after some thought, "I can stand on the sidewalk and flip a coin whenever someone approaches. If heads, I'll look up; if tails, I'll look straight ahead. I'll repeat the experiment a few thousand times. If the proportion of pedestrians who looked up within 10 seconds after I did is greater than the proportion of pedestrians who looked up when I didn't, I will conclude that my looking up has a causal effect on other people's looking up. By the way, I may hire an assistant to record what people do while I'm looking up." After conducting this study, you found that 55% of pedestrians looked up when you looked up but only 1% looked up when you looked straight ahead.

Your solution to our challenge was to conduct a randomized experiment. It was an experiment because the investigator (you) carried out the action of interest (looking up), and it was randomized because the decision to act on any study subject (pedestrian) was made by a random device (coin flipping). Not all experiments are randomized. For example, you could have looked up when a man approached and looked straight ahead when a woman did. Then the assignment of the action would have followed a deterministic rule (up for man, straight for woman) rather than a random mechanism. However, your findings would not have been nearly as convincing if you had conducted a nonrandomized experiment. If your action had been determined by the pedestrian's sex, critics could argue that the "looking up" behavior of men and women differs (women may look up less often than do men after you look up) and thus your study compared essentially "noncomparable" groups of people. This chapter describes why randomization results in convincing causal inferences.

2.1 Randomization

Neyman (1923) applied counterfactual theory to the estimation of causal effects via randomized experiments.

In a real world study we will not know both of Zeus's potential outcomes $Y^{a=1}$ under treatment and $Y^{a=0}$ under no treatment. Rather, we can only know his observed outcome Y under the treatment value A that he happened to receive. Table 2.1 summarizes the available information for our population of 20 individuals. Only one of the two counterfactual outcomes is known for each individual: the one corresponding to the treatment level that he actually received. The data are missing for the other counterfactual outcomes. As we discussed in the previous chapter, this missing data creates a problem because it appears that we need the value of both counterfactual outcomes to compute effect measures. The data in Table 2.1 are only good to compute association measures.

Randomized experiments, like any other real world study, generate data with missing values of the counterfactual outcomes as shown in Table 2.1. However, randomization ensures that those missing values occurred by chance. As a result, effect measures can be computed —or, more rigorously, consistently estimated—in randomized experiments despite the missing data. Let us be more precise.

Suppose that the population represented by a diamond in Figure 1.1 was

Table 2.1

	A	Y	Y^0	Y^1
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaestus	1	1	?	1
Aphrodite	1	1	?	1
Polyphemus	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

near-infinite, and that we flipped a coin for each individual in such population. We assigned the individual to the white group if the coin turned tails, and to the grey group if it turned heads. Note this was not a fair coin because the probability of heads was less than 50%—fewer people ended up in the grey group than in the white group. Next we asked our research assistants to administer the treatment of interest ($A = 1$), to individuals in the white group and a placebo ($A = 0$) to those in the grey group. Five days later, at the end of the study, we computed the mortality risks in each group, $\Pr[Y = 1|A = 1] = 0.3$ and $\Pr[Y = 1|A = 0] = 0.6$. The associational risk ratio was $0.3/0.6 = 0.5$ and the associational risk difference was $0.3 - 0.6 = -0.3$. We will assume that this was an *ideal randomized experiment* in all other respects: no loss to follow-up, full adherence to the assigned treatment over the duration of the study, a single version of treatment, and double blind assignment (see Chapter 9). Ideal randomized experiments are unrealistic but useful to introduce some key concepts for causal inference. Later in this book we consider more realistic randomized experiments.

Now imagine what would have happened if the research assistants had misinterpreted our instructions and had treated the grey group rather than the white group. Say we learned of the misunderstanding after the study finished. How does this reversal of treatment status affect our conclusions? Not at all. We would still find that the risk in the treated (now the grey group) $\Pr[Y = 1|A = 1]$ is 0.3 and the risk in the untreated (now the white group) $\Pr[Y = 1|A = 0]$ is 0.6. The association measure would not change. Because individuals were randomly assigned to white and grey groups, the proportion of deaths among the exposed, $\Pr[Y = 1|A = 1]$ is expected to be the same whether individuals in the white group received the treatment and individuals in the grey group received placebo, or vice versa. When group membership is randomized, which particular group received the treatment is irrelevant for the value of $\Pr[Y = 1|A = 1]$. The same reasoning applies to $\Pr[Y = 1|A = 0]$, of course. Formally, we say that groups are exchangeable.

Exchangeability means that the risk of death in the white group would have been the same as the risk of death in the grey group had individuals in the white group received the treatment given to those in the grey group. That is, the risk under the potential treatment value a among the treated, $\Pr[Y^a = 1|A = 1]$, equals the risk under the potential treatment value a among the untreated, $\Pr[Y^a = 1|A = 0]$, for both $a = 0$ and $a = 1$. An obvious consequence of these (conditional) risks being equal in all subsets defined by treatment status in the population is that they must be equal to the (marginal) risk under treatment value a in the whole population: $\Pr[Y^a = 1|A = 1] = \Pr[Y^a = 1|A = 0] = \Pr[Y^a = 1]$. Because the counterfactual risk under treatment value a is the same in both groups $A = 1$ and $A = 0$, we say that the actual treatment A does not predict the counterfactual outcome Y^a. Equivalently, exchangeability means that the counterfactual outcome and the actual treatment are independent, or $Y^a \perp\!\!\!\perp A$, for all values a. Randomization is so highly valued because it is expected to produce exchangeability. When the treated and the untreated are exchangeable, we sometimes say that treatment is exogenous, and thus *exogeneity* is commonly used as a synonym for exchangeability.

Exchangeability:
$Y^a \perp\!\!\!\perp A$ for all a. See also Technical Point 2.1 for other versions of exchangeability.

The previous paragraph argues that, in the presence of exchangeability, the counterfactual risk under treatment in the white part of the population would equal the counterfactual risk under treatment in the entire population. But the risk under treatment in the white group is not counterfactual at all because the white group was actually treated! Therefore our ideal randomized experiment allows us to compute the counterfactual risk under treatment in the population

Technical Point 2.1

Full exchangeability and mean exchangeability. Randomization makes the Y^a jointly independent of A which implies, but is not implied by, exchangeability $Y^a \perp\!\!\!\perp A$ for each a. Formally, let $\mathcal{A} = \{a, a', a'', ...\}$ denote the set of all treatment values present in the population, and $Y^{\mathcal{A}} = \left\{Y^a, Y^{a'}, Y^{a''}, ...\right\}$ the set of all counterfactual outcomes. Randomization makes $Y^{\mathcal{A}} \perp\!\!\!\perp A$. We refer to this joint independence as *full exchangeability*. For a dichotomous treatment, $\mathcal{A} = \{0, 1\}$ and full exchangeability is $\left(Y^{a=1}, Y^{a=0}\right) \perp\!\!\!\perp A$.

For a dichotomous outcome and treatment, exchangeability $Y^a \perp\!\!\!\perp A$ can also be written as $\Pr\left[Y^a = 1|A = 1\right] = \Pr\left[Y^a = 1|A = 0\right]$ or, equivalently, as $\mathrm{E}[Y^a|A = 1] = \mathrm{E}[Y^a|A = 0]$ for all a. We refer to the last equality as *mean exchangeability*. For a continuous outcome, exchangeability $Y^a \perp\!\!\!\perp A$ implies mean exchangeability $\mathrm{E}[Y^a|A = a'] = \mathrm{E}[Y^a]$, but mean exchangeability does not imply exchangeability because distributional parameters other than the mean (e.g., variance) may not be independent of treatment.

Neither full exchangeability $Y^{\mathcal{A}} \perp\!\!\!\perp A$ nor exchangeability $Y^a \perp\!\!\!\perp A$ are required to prove that $\mathrm{E}[Y^a] = \mathrm{E}[Y|A = a]$. Mean exchangeability is sufficient. As sketched in the main text, the proof has two steps. First, $\mathrm{E}[Y|A = a] = \mathrm{E}[Y^a|A = a]$ by consistency. Second, $\mathrm{E}[Y^a|A = a] = \mathrm{E}[Y^a]$ by mean exchangeability. Because exchangeability and mean exchangeability are identical concepts for the dichotomous outcomes used in this chapter, we use the shorter term "exchangeability" throughout.

$\Pr\left[Y^{a=1} = 1\right]$ because it is equal to the risk in the treated $\Pr\left[Y = 1|A = 1\right] = 0.3$. That is, the risk in the treated (the white part of the diamond) is the same as the risk if everybody had been treated (and thus the diamond had been entirely white). Of course, the same rationale applies to the untreated: the counterfactual risk under no treatment in the population $\Pr\left[Y^{a=0} = 1\right]$ equals the risk in the untreated $\Pr\left[Y = 1|A = 0\right] = 0.6$. The causal risk ratio is 0.5 and the causal risk difference is -0.3. In ideal randomized experiments, association *is* causation.

Here is another explanation for exchangeability $Y^a \perp\!\!\!\perp A$ in a randomized experiment. The counterfactual outcome Y^a, like one's genetic make-up, can be thought of as a fixed characteristic of a person existing before the treatment A was randomly assigned. This is because Y^a encodes what would have been one's outcome if assigned to treament a and thus does not depend on the treatment you later receive. Because treatment A was randomized, it is independent of both your genes and Y^a. The difference between Y^a and your genetic make-up is that, even conceptually, you can only learn the value of Y^a after treatment is given and then only if one's treatment A is equal to a.

Caution:
$Y^a \perp\!\!\!\perp A$ is different from $Y \perp\!\!\!\perp A$.

Before proceeding, please make sure you understand the difference between $Y^a \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A$. Exchangeability $Y^a \perp\!\!\!\perp A$ is defined as independence between the counterfactual outcome and the observed treatment. Again, this means that the treated and the untreated would have experienced the same risk of death if they had received the same treatment level (either $a = 0$ or $a = 1$). But independence between the counterfactual outcome and the observed treatment $Y^a \perp\!\!\!\perp A$ does not imply independence between the observed outcome and the observed treatment $Y \perp\!\!\!\perp A$. For example, in a randomized experiment in which exchangeability $Y^a \perp\!\!\!\perp A$ holds and the treatment has a causal effect on the outcome, then $Y \perp\!\!\!\perp A$ does not hold because the treatment is associated with the observed outcome.

Suppose there is a causal effect on some individuals so that $Y^{a=1} \neq Y^{a=0}$. Since $Y = Y^A$, then Y^a with a evaluated at the observed treatment A is the observed Y^A, which depends on A, and thus will not be independent of A.

Does exchangeability hold in our heart transplant study of Table 2.1? To answer this question we would need to check whether $Y^a \perp\!\!\!\perp A$ holds for $a = 0$ and for $a = 1$. Take $a = 0$ first. Suppose the counterfactual data in Table 1.1 are available to us. We can then compute the risk of death under no treatment

Fine Point 2.1

Crossover experiments. Suppose we want to estimate the individual causal effect of lightning bolt use A on Zeus's blood pressure Y. We define the counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$ to be 1 if Zeus's blood pressure is temporarily elevated after calling or not calling a lightning strike, respectively. Suppose we convinced Zeus to use his lightning bolt only when suggested by us. Yesterday morning we asked Zeus to call a lightning strike ($a = 1$). His blood pressure was elevated after doing so. This morning we asked Zeus to refrain from using his lightning bolt ($a = 0$). His blood pressure did not increase. We have conducted a *crossover experiment* in which an individual's outcome is sequentially observed under two treatment values. One might argue that, because we have observed both of Zeus's counterfactual outcomes $Y^{a=1} = 1$ and $Y^{a=0} = 0$, using a lightning bolt has a causal effect on Zeus's blood pressure. However, this argument is generally incorrect unless the very strong assumptions i)–iii) given in the next paragraph are true.

In crossover experiments, individuals are observed during two or more periods, say $t = 0$ and $t = 1$. An individual i receives a different treatment value A_{it} in each period t. Let $Y_{i1}^{a_0 a_1}$ be the (deterministic) counterfactual outcome at $t = 1$ for individual i if treated with a_1 at $t = 1$ and a_0 at $t = 0$. Let $Y_{i0}^{a_0}$ be defined similarly for $t = 0$. The individual causal effect $Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$ can be identified if the following three conditions hold: i) no carryover effect of treatment: $Y_{it=1}^{a_0,a_1} = Y_{it=1}^{a_1}$, ii) the individual causal effect does not depend on time: $Y_{it}^{a_t=1} - Y_{it}^{a_t=0} = \alpha_i$ for $t = 0, 1$, and iii) the counterfactual outcome under no treatment does not depend on time: $Y_{it}^{a_t=0} = \beta_i$ for $t = 0, 1$. Under these conditions, if the individual is treated at time 1 ($A_{i1} = 1$) but not time 0 ($A_{i0} = 0$) then, by consistency, $Y_{i1} - Y_{i0}$ is the individual causal effect because $Y_{i1} - Y_{i0} = Y_{i1}^{a_1=1} - Y_{i0}^{a_0=0} = Y_{i1}^{a_1=1} - Y_{i1}^{a_1=0} + Y_{i1}^{a_1=0} - Y_{i0}^{a_0=0} = \alpha_i + \beta_i - \beta_i = \alpha_i$. Similarly if $A_{i1} = 0$ and $A_{i0} = 1$, $Y_{i0} - Y_{i1} = \alpha_i$ is the individual level causal effect.

Condition (i) implies that the outcome $Y_{it}^{a_t}$ has an abrupt onset that completely resolves by the next time period. Hence, crossover experiments cannot be used to study the effect of heart transplant, an irreversible action, on death, an irreversible outcome. See also Fine Point 3.2.

Reminder: Our discussion of randomized experiments refers to population or average causal effects because individual causal effects cannot generally be identified. See Fine Point 2.1.

$\Pr\left[Y^{a=0} = 1|A = 1\right] = 7/13$ in the 13 treated individuals and the risk of death under no treatment $\Pr\left[Y^{a=0} = 1|A = 0\right] = 3/7$ in the 7 untreated individuals. Since the risk of death under no treatment is greater in the treated than in the untreated individuals, i.e., $7/13 > 3/7$, we conclude that the treated have a worse prognosis than the untreated, i.e., that the treated and the untreated are not exchangeable. Mathematically, we have proven that exchangeability $Y^a \perp\!\!\!\perp A$ does not hold for $a = 0$. (You can check that it does not hold for $a = 1$ either.) Thus the answer to the question that opened this paragraph is 'No'.

But only the observed data in Table 2.1, not the counterfactual data in Table 1.1, are available in the real world. Since Table 2.1 is insufficient to compute counterfactual risks like the risk under no treatment in the treated $\Pr\left[Y^{a=0} = 1|A = 1\right]$, we are generally unable to determine whether exchangeability holds in our study. However, suppose for a moment, that we actually had access to Table 1.1 and determined that exchangeability does not hold in our heart transplant study. Can we then conclude that our study is not a randomized experiment? No, for two reasons. First, as you are probably already thinking, a twenty-person study is too small to reach definite conclusions. Random fluctuations arising from sampling variability could explain almost anything. We will discuss random variability in Chapter 10. Until then, let us assume that each individual in our population represents 1 billion individuals that are identical to him or her. Second, it is still possible that a study is a randomized experiment even if exchangeability does not hold in infinite samples. However, unlike the type of randomized experiment described in this section, it would need to be a randomized experiment in which investigators use more than one coin to randomly assign treatment. The next section describes randomized experiments with more than one coin.

2.2 Conditional randomization

Table 2.2

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Polyphemus	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Table 2.2 shows the data from our heart transplant randomized study. Besides data on treatment A (1 if the individual received a transplant, 0 otherwise) and outcome Y (1 if the individual died, 0 otherwise), Table 2.2 also contains data on the prognostic factor L (1 if the individual was in critical condition, 0 otherwise), which we measured before treatment was assigned. We now consider two mutually exclusive study designs and discuss whether the data in Table 2.2 could have arisen from either of them.

In design 1 we would have randomly selected 65% of the individuals in the population and transplanted a new heart to each of the selected individuals. That would explain why 13 out of 20 individuals were treated. In design 2 we would have classified all individuals as being in either critical ($L = 1$) or noncritical ($L = 0$) condition. Then we would have randomly selected 75% of the individuals in critical condition and 50% of those in noncritical condition, and transplanted a new heart to each of the selected individuals. That would explain why 9 out of 12 individuals in critical condition, and 4 out of 8 individuals in noncritical condition, were treated.

Both designs are randomized experiments. Design 1 is precisely the type of randomized experiment described in Section 2.1. Under this design, we would use a single coin to assign treatment to all individuals (e.g., treated if tails, untreated if heads): a loaded coin with probability 0.65 of turning tails, thus resulting in 65% of the individuals receiving treatment. Under design 2 we would not use a single coin for all individuals. Rather, we would use a coin with a 0.75 chance of turning tails for individuals in critical condition, and another coin with a 0.50 chance of turning tails for individuals in noncritical condition. We refer to design 2 experiments as *conditionally randomized experiments* because we use several randomization probabilities that depend (are conditional) on the values of the variable L. We refer to design 1 experiments as *marginally randomized experiments* because we use a single unconditional (marginal) randomization probability that is common to all individuals.

As discussed in the previous section, a marginally randomized experiment is expected to result in exchangeability of the treated and the untreated:

$$\Pr[Y^a = 1|A = 1] = \Pr[Y^a = 1|A = 0] \quad \text{or} \quad Y^a \perp\!\!\!\perp A \quad \text{for all } a.$$

In contrast, a conditionally randomized experiment will not generally result in exchangeability of the treated and the untreated because, by design, each group may have a different proportion of individuals with bad prognosis.

Thus the data in Table 2.2 could not have arisen from a marginally randomized experiment because 69% treated versus 43% untreated individuals were in critical condition. This imbalance indicates that the risk of death in the treated, had they remained untreated, would have been higher than the risk of death in the untreated. That is, treatment A predicts the counterfactual risk of death under no treatment, and exchangeability $Y^a \perp\!\!\!\perp A$ does not hold. Since our study was a randomized experiment, you can safely conclude that the study was a randomized experiment with randomization conditional on L.

Our conditionally randomized experiment is simply the combination of two separate marginally randomized experiments: one conducted in the subset of individuals in critical condition ($L = 1$), the other in the subset of individuals in noncritical condition ($L = 0$). Consider first the randomized experiment being conducted in the subset of individuals in critical condition. In this subset, the treated and the untreated are exchangeable. Formally, the counterfactual mortality risk under each treatment value a is the same among the treated

and the untreated given that they all were in critical condition at the time of treatment assignment. That is,

$$\Pr[Y^a = 1|A = 1, L = 1] = \Pr[Y^a = 1|A = 0, L = 1] \text{ or } Y^a \perp\!\!\!\perp A|L = 1 \text{ for all } a,$$

where $Y^a \perp\!\!\!\perp A|L = 1$ means Y^a and A are independent given $L = 1$. Similarly, randomization also ensures that the treated and the untreated are exchangeable in the subset of individuals that were in noncritical condition, i.e., $Y^a \perp\!\!\!\perp A|L = 0$. When $Y^a \perp\!\!\!\perp A|L = l$ holds for all values l we simply write $Y^a \perp\!\!\!\perp A|L$. Thus, although conditional randomization does not guarantee unconditional (or marginal) exchangeability $Y^a \perp\!\!\!\perp A$, it guarantees *conditional exchangeability* $Y^a \perp\!\!\!\perp A|L$ within levels of the variable L. In summary, marginal randomization (design 1) produces both marginal exchangeability and conditional exchangeability, whereas conditional randomization (design 2) produces only conditional exchangeability.

Conditional exchangeability: $Y^a \perp\!\!\!\perp A|L$ for all a

We know how to compute effect measures under marginal exchangeability: the causal risk ratio $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ equals the associational risk ratio $\Pr[Y = 1|A = 1] / \Pr[Y = 1|A = 0]$ in marginally randomized experiments because exchangeability ensures that the counterfactual risk under treatment level a, $\Pr[Y^a = 1]$, equals the observed risk among those who received treatment level a, $\Pr[Y = 1|A = a]$. Thus, if the data in Table 2.2 had been collected during a marginally randomized experiment, the causal risk ratio would be readily calculated from the data on A and Y as $\frac{7/13}{3/7} = 1.26$. The question is how to compute the causal risk ratios in a conditionally randomized experiment. Remember that a conditionally randomized experiment is simply the combination of two (or more) separate marginally randomized experiments conducted in different subsets of the population $L = 1$ and $L = 0$. Thus we have two options.

If $A = 1$, $Y^{a=0}$ is missing; if $A = 0$, $Y^{a=1}$ is missing. Data are missing completely at random (MCAR) if $\Pr[A = a|L, Y^{a=1}, Y^{a=0}] = \Pr[A = a]$, which holds in a marginally randomized experiment. Data are missing at random (MAR) if the probability of $A = a$ conditional on the full data $(L, Y^{a=1}, Y^{a=0})$ only depends on the data that would be observed (L, Y^a) if $A = a$. In fact, MAR implies $\Pr[A = a|L, Y^{a=1}, Y^{a=0}] = \Pr[A = a|L]$, which holds in a conditionally randomized experiment because, by MAR, $\Pr[A = 1|L, Y^{a=1}, Y^{a=0}]$ cannot depend on $Y^{a=0}$ and $1 - \Pr[A = 1|L, Y^{a=1}, Y^{a=0}] = \Pr[A = 0|L, Y^{a=1}, Y^{a=0}]$ cannot depend on $Y^{a=1}$. The terms MCAR, MAR, and MNAR (missing not at random) were introduced by Rubin (1976) and Marini, Olsen, and Rubin (1980).

First, we compute the average causal effect in each of these subsets or strata of the population. Because association is causation within each subset, the stratum-specific causal risk ratio $\Pr[Y^{a=1} = 1|L = 1] / \Pr[Y^{a=0} = 1|L = 1]$ among people in critical condition is equal to the stratum-specific associational risk ratio $\Pr[Y = 1|L = 1, A = 1] / \Pr[Y = 1|L = 1, A = 0]$ among people in critical condition. And analogously for $L = 0$. We refer to this method to compute stratum-specific causal effects as *stratification*. Note that the stratum-specific causal risk ratio in the subset $L = 1$ may differ from the causal risk ratio in $L = 0$. In that case, we say that the effect of treatment is modified by L, or that there is *effect modification* by L or that there is *treatment effect heterogeneity* across levels of L. Stratification and effect modification are discussed in more detail in Chapter 4.

Second, we compute the average causal effect $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ in the entire population, as we have been doing so far. Whether our principal interest lies in the stratum-specific average causal effects versus the average causal effect in the entire population depends on practical and theoretical considerations discussed in detail in Chapter 4 and in Part III. As one example, you may be interested in the average causal effect in the entire population, rather than in the stratum-specific average causal effects, if you do not expect to have information on L for future individuals (e.g., the variable L is expensive to measure) and thus your decision to treat cannot depend on the value of L. Until Chapter 4, we will restrict our attention to the average causal effect in the entire population. The next two sections describe how to use data from conditionally randomized experiments to compute the average causal effect in the entire population. See also Fine Point 2.2 for a discussion of risk periods.

Fine Point 2.2

Risk periods. We have defined a risk as the proportion of individuals who develop the outcome of interest during a particular period. For example, the 5-day mortality risk in the treated $\Pr[Y = 1|A = 1]$ is the proportion of treated individuals who died during the first five days of follow-up. Throughout the book we often specify the period when the risk is first defined (e.g., 5 days) and, for conciseness, omit it later. That is, we may just say "the mortality risk" rather than "the five-day mortality risk."

The following example highlights the importance of specifying the risk period. Suppose a randomized experiment was conducted to quantify the causal effect of antibiotic therapy on mortality among elderly humans infected with the plague bacteria. An investigator analyzes the data and concludes that the causal risk ratio is 0.05, i.e., on average antibiotics decrease mortality by 95%. A second investigator also analyzes the data but concludes that the causal risk ratio is 1, i.e., antibiotics have a null average causal effect on mortality. Both investigators are correct. The first investigator computed the ratio of 1-year risks, whereas the second investigator computed the ratio of 100-year risks. The 100-year risk was of course 1 regardless of whether individuals received the treatment. When we say that a treatment has a causal effect on mortality, we mean that death is delayed, not prevented, by the treatment.

2.3 Standardization

Our heart transplant study is a conditionally randomized experiment: the investigators used a random procedure to assign hearts ($A = 1$) with probability 50% to the 8 individuals in noncritical condition ($L = 0$), and with probability 75% to the 12 individuals in critical condition ($L = 1$). First, let us focus on the 8 individuals—remember, they are really the average representatives of 8 billion individuals—in noncritical condition. In this group, the risk of death among the treated is $\Pr[Y = 1|L = 0, A = 1] = \frac{1}{4}$, and the risk of death among the untreated is $\Pr[Y = 1|L = 0, A = 0] = \frac{1}{4}$. Because treatment was randomly assigned to individuals in the group $L = 0$, i.e., $Y^a \perp\!\!\!\perp A|L = 0$, the observed risks are equal to the counterfactual risks. That is, in the group $L = 0$, the risk in the treated equals the risk if everybody had been treated, $\Pr[Y = 1|L = 0, A = 1] = \Pr\left[Y^{a=1} = 1|L = 0\right]$, and the risk in the untreated equals the risk if everybody had been untreated, $\Pr[Y = 1|L = 0, A = 0] = \Pr\left[Y^{a=0} = 1|L = 0\right]$. Following a similar reasoning, we can conclude that the observed risks equal the counterfactual risks in the group of 12 individuals in critical condition, i.e., $\Pr[Y = 1|L = 1, A = 1] = \Pr\left[Y^{a=1} = 1|L = 1\right] = \frac{2}{3}$, and $\Pr[Y = 1|L = 1, A = 0] = \Pr\left[Y^{a=0} = 1|L = 1\right] = \frac{2}{3}$.

Suppose now that our goal is to compute the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. The numerator of the causal risk ratio is the risk if all 20 individuals in the population had been treated. From the previous paragraph, we know that the risk if all individuals had been treated is $\frac{1}{4}$ in the 8 individuals with $L = 0$ and $\frac{2}{3}$ in the 12 individuals with $L = 1$. Therefore the risk if all 20 individuals in the population had been treated will be a weighted average of $\frac{1}{4}$ and $\frac{2}{3}$ in which each group receives a weight proportional to its size. Since 40% of the individuals (8) are in group $L = 0$ and 60% of the individuals (12) are in group $L = 1$, the weighted average is $\frac{1}{4} \times 0.4 + \frac{2}{3} \times 0.6 = 0.5$. Thus the risk if everybody had been treated $\Pr[Y^{a=1} = 1]$ is equal to 0.5. By following the same reasoning we can calculate that the risk if nobody had been treated $\Pr[Y^{a=0} = 1]$ is also equal to 0.5. The causal risk ratio is then $0.5/0.5 = 1$.

More formally, the marginal counterfactual risk $\Pr[Y^a = 1]$ is the weighted average of the stratum-specific risks $\Pr[Y^a = 1|L = 0]$ and $\Pr[Y^a = 1|L = 1]$ with weights equal to the proportion of individuals in the population with

$L = 0$ and $L = 1$, respectively. That is, $\Pr[Y^a = 1] = \Pr[Y^a = 1|L = 0]\Pr[L = 0] + \Pr[Y^a = 1|L = 1]\Pr[L = 1]$. Or, using a more compact notation, $\Pr[Y^a = 1] = \sum_l \Pr[Y^a = 1|L = l]\Pr[L = l]$, where $\sum_l$ means sum over all values l that occur in the population. Under conditional exchangeability, we can replace the counterfactual risk $\Pr[Y^a = 1|L = l]$ by the observed risk $\Pr[Y = 1|L = l, A = a]$ in the expression above. That is, $\Pr[Y^a = 1] = \sum_l \Pr[Y = 1|L = l, A = a]\Pr[L = l]$. The left-hand side of this equality is an unobserved counterfactual risk whereas the right-hand side includes observed quantities only, which can be computed using data on L, A, and Y. When, as here, a counterfactual quantity can be expressed as a function of the distribution (i.e., the probabilities) of the observed data, we say that the counterfactual quantity is identified (or identifiable); otherwise, we say it is unidentified.

Standardized mean
$\sum_l \mathrm{E}[Y|L = l, A = a]$
$\times \Pr[L = l]$

This method is known in epidemiology, demography, and other disciplines as *standardization*. For example, the numerator $\sum_l \Pr[Y = 1|L = l, A = 1]\Pr[L = l]$ of the causal risk ratio is the standardized risk in the treated using the population as the standard. Under conditional exchangeability, this standardized risk can be interpreted as the (counterfactual) risk that would have been observed had all the individuals in the population been treated.

The standardized risks in the treated and the untreated are equal to the counterfactual risks under treatment and no treatment, respectively. Therefore, the causal risk ratio $\dfrac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]}$ can be computed by standardization as $\dfrac{\sum_l \Pr[Y = 1|L = l, A = 1]\Pr[L = l]}{\sum_l \Pr[Y = 1|L = l, A = 0]\Pr[L = l]}$.

2.4 Inverse probability weighting

In the previous section we computed the causal risk ratio in a conditionally randomized experiment via standardization. In this section we compute this causal risk ratio via inverse probability weighting. The data in Table 2.2 can be displayed as a tree in which all 20 individuals start at the left and progress over time towards the right, as in Figure 2.1. The leftmost circle of the tree contains its first branching: 8 individuals were in noncritical condition ($L = 0$) and 12 in critical condition ($L = 1$). The numbers in parentheses are the probabilities of being in noncritical, $\Pr[L = 0] = 8/20 = 0.4$, or critical, $\Pr[L = 1] = 12/20 = 0.6$, condition. Let us follow, e.g., the branch $L = 0$. Of the 8 individuals in this branch, 4 were untreated ($A = 0$) and 4 were treated ($A = 1$). The conditional probability of being untreated is $\Pr[A = 0|L = 0] = 4/8 = 0.5$, as shown in parentheses. The conditional probability of being treated $\Pr[A = 1|L = 0]$ is 0.5 too. The upper right circle represents that, of the 4 individuals in the branch ($L = 0, A = 0$), 3 survived ($Y = 0$) and 1 died ($Y = 1$). That is, $\Pr[Y = 0|L = 0, A = 0] = 3/4$ and $\Pr[Y = 1|L = 0, A = 0] = 1/4$. The other branches of the tree are interpreted analogously. The circles contain the bifurcations defined by non-treatment variables. We now use this tree to compute the causal risk ratio.

Figure 2.1 is an example of a fully randomized causally interpreted structured tree graph or FR-CISTG (Robins 1986, 1987) representation of a conditionally randomized experiment. Did we win the prize for the worst acronym ever?

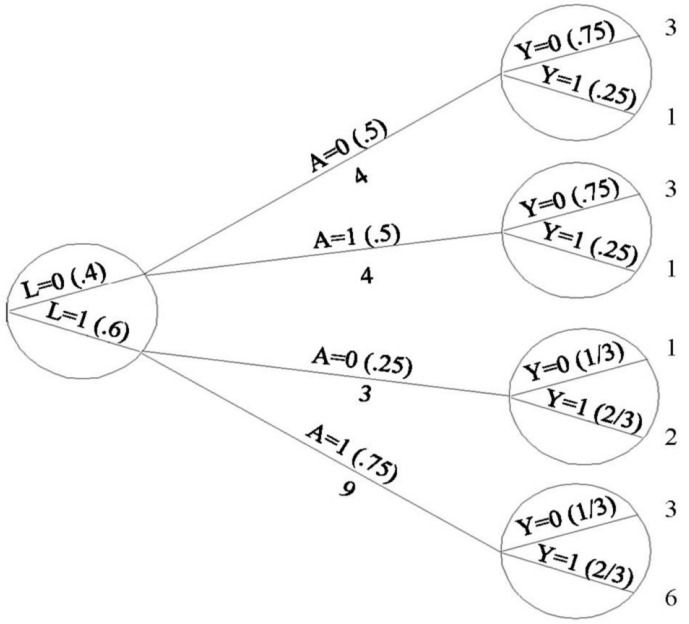

Figure 2.1

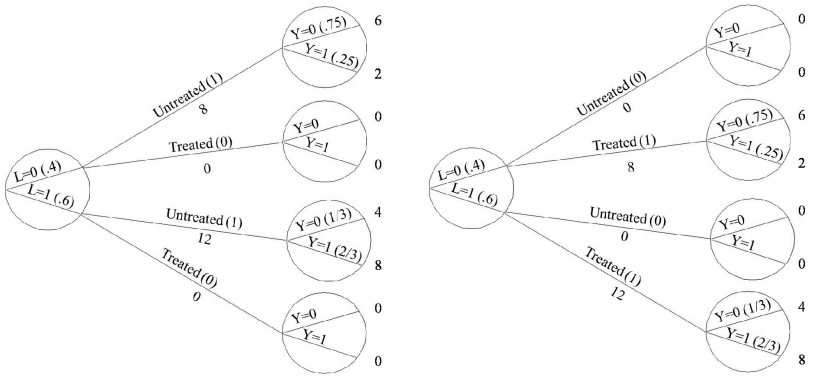

Figure 2.2

The denominator of the causal risk ratio, $\Pr[Y^{a=0} = 1]$, is the counterfactual risk of death had everybody in the population remained untreated. Let us calculate this risk. In Figure 2.1, 4 out of 8 individuals with $L = 0$ were untreated, and 1 of them died. How many deaths would have occurred had the 8 individuals with $L = 0$ remained untreated? Two deaths, because if 8 individuals rather than 4 individuals had remained untreated, then 2 deaths rather than 1 death would have been observed. If the number of individuals is multiplied times 2, then the number of deaths is also doubled. In Figure 2.1, 3 out of 12 individuals with $L = 1$ were untreated, and 2 of them died. How many deaths would have occurred had the 12 individuals with $L = 1$ remained untreated? Eight deaths, or 2 deaths times 4, because 12 is 3×4. That is, if all $8 + 12 = 20$ individuals in the population had been untreated, then $2 + 8 = 10$ would have died. The denominator of the causal risk ratio, $\Pr[Y^{a=0} = 1]$, is $10/20 = 0.5$. The first tree in Figure 2.2 shows the population had everybody

remained untreated. Of course, these calculations rely on the condition that treated individuals with $L = 0$, had they remained untreated, would have had the same probability of death as those who actually remained untreated. This condition is precisely exchangeability given $L = 0$.

The numerator of the causal risk ratio $\Pr[Y^{a=1} = 1]$ is the counterfactual risk of death had everybody in the population been treated. Reasoning as in the previous paragraph, this risk is calculated to be also $10/20 = 0.5$, under exchangeability given $L = 1$. The second tree in Figure 2.2 shows the population had everybody been treated. Combining the results from this and the previous paragraph, the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$ is equal to $0.5/0.5 = 1$. We are done.

Let us examine how this method works. The two trees in Figure 2.2 are a simulation of what would have happened had all individuals in the population been untreated and treated, respectively. These simulations are correct under conditional exchangeability. Both simulations can be pooled to create a hypothetical population in which every individual appears as a treated and as an untreated individual. This hypothetical population, twice as large as the original population, is known as the *pseudo-population*. Figure 2.3 shows the entire pseudo-population. Under conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ in the original population, the treated and the untreated are (unconditionally) exchangeable in the pseudo-population because the L is independent of A. That is, the associational risk ratio in the pseudo-population is equal to the causal risk ratio in both the pseudo-population and the original population.

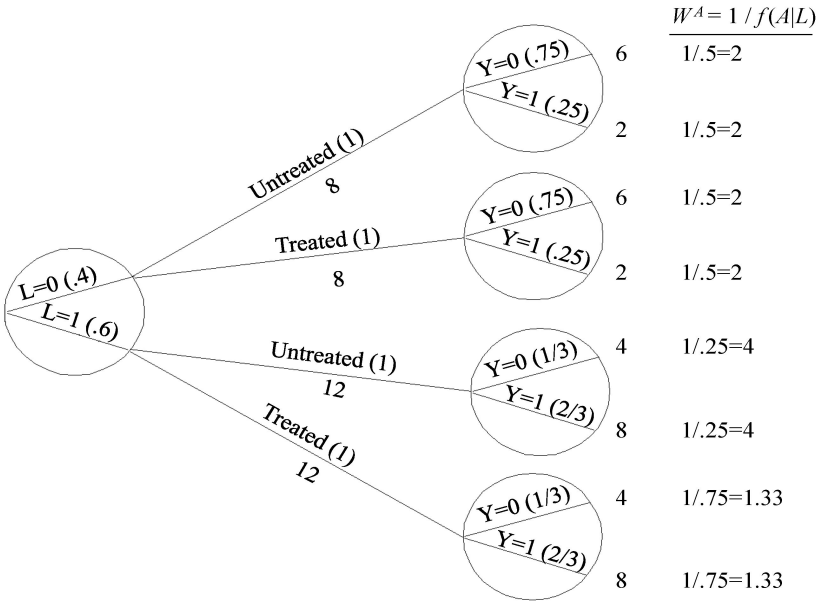

Figure 2.3

IP weighted estimators were proposed by Horvitz and Thompson (1952) for surveys in which subjects are sampled with unequal probabilities. See Technical Point 12.1.

This method is known as *inverse probability (IP) weighting*. To see why, let us look at, say, the 4 untreated individuals with $L = 0$ in the population of Figure 2.1. These individuals are used to create 8 members of the pseudo-population of Figure 2.3. That is, each of them receives a weight of 2, which is equal to $1/0.5$. Figure 2.1 shows that 0.5 is the conditional probability of staying untreated given $L = 0$. Similarly, the 9 treated individuals with $L = 1$

Technical Point 2.2

Formal definition of IP weights. An individual's IP weight depends on the individual's values of treatment A and covariate L. For example, a treated individual with $L = l$ receives the weight $1/\Pr[A = 1|L = l]$, whereas an untreated individual with $L = l'$ receives the weight $1/\Pr[A = 0|L = l']$. We can express these weights using a single expression for all individuals—regardless of their individual treatment and covariate values—by using the probability density function (pdf) of A rather than the probability of A. The conditional pdf of A given L evaluated at the values a and l is represented by $f_{A|L}[a|l]$, or simply as $f[a|l]$. For discrete variables A and L, $f[a|l]$ is the conditional probability $\Pr[A = a|L = l]$. In a conditionally randomized experiment, $f[a|l]$ is positive for all l such that $\Pr[L = l]$ is nonzero.

Since the denominator of the weight for each individual is the conditional density evaluated at the individual's own values of A and L, it can be expressed as the conditional density evaluated at the random arguments A and L (as opposed to the fixed arguments a and l), that is, as $f[A|L]$. This notation, which appeared in Figure 2.3, is used to define the IP weights $W^A = 1/f[A|L]$. It is needed to have a unified notation for the weights because $\Pr[A = A|L = L]$ is tautologically equal to 1 and thus not considered proper notation.

As explained in the main text, the mean of the outcome in the pseudo-population $\mathrm{E}_{ps}[Y|A = a]$ equals the IP weighted mean of the outcome in the population, $\mathrm{E}[Y\,\mathrm{I}(A = a)/\Pr(A = a|L)]$, where $\mathrm{I}(A = a)$ is 1 when $A = a$ and 0 otherwise. A proof follows:

$\mathrm{E}_{ps}[Y|A = a] = \mathrm{E}_{ps}[Y\,\mathrm{I}(A = a)]/\mathrm{E}_{ps}[\mathrm{I}(A = a)]$ (by the laws of probability)
$= \mathrm{E}\left[W^A Y\,\mathrm{I}(A = a)\right]/\mathrm{E}\left[\mathrm{I}(A = a)\,W^A\right]$ (by definition of E_{ps})
$= \mathrm{E}[Y\,\mathrm{I}(A = a)/\Pr(A = a|L)]/\mathrm{E}[\mathrm{I}(A = a)/\Pr(A = a|L)]$ (because $\mathrm{I}(A = a)/f(A|L) = \mathrm{I}(A = a)/f(a|L)$)
$= \mathrm{E}[Y\,\mathrm{I}(A = a)/\Pr(A = a|L)]$ (because $\mathrm{E}[\mathrm{I}(A = a)/\Pr(A = a|L)\,|L] = 1$).

IP weight: $W^A = 1/f[A|L]$

in Figure 2.1 are used to create 12 members of the pseudo-population. That is, each of them receives a weight of $1.33 = 1/0.75$. Figure 2.1 shows that 0.75 is the conditional probability of being treated given $L = 1$. Informally, the pseudo-population is created by weighting each individual in the population by the inverse of the conditional probability of receiving the treatment level that she indeed received. These IP weights are shown in Figure 2.3.

IP weighting yielded the same result as standardization—causal risk ratio equal to 1— in our example above. This is no coincidence: standardization and IP weighting are mathematically equivalent (see Technical Point 2.3). In fact, both standardization and IP weighting can be viewed as procedures to build a new tree in which all individuals receive treatment a. Each method uses a different set of the probabilities to build the counterfactual tree: IP weighting uses the conditional probability of treatment A given the covariate L (as shown in Figure 2.1), standardization uses the probability of the covariate L and the conditional probability of outcome Y given A and L.

Because both standardization and IP weighting simulate what would have been observed if the variable (or variables in the vector) L had not been used to decide the probability of treatment, we often say that these methods *adjust for* L. In a slight abuse of language we sometimes say that these methods *control for* L, but this "analytic control" is quite different from the "physical control" in a randomized experiment. Standardization and IP weighting can be generalized to conditionally randomized studies with continuous outcomes (see Technical Point 2.3).

Why not finish this book here? We have a study design (an ideal randomized experiment) that, when combined with the appropriate analytic method (standardization or IP weighting), allows us to compute average causal effects. Unfortunately, randomized experiments are often unethical, impractical, or untimely. For example, it is questionable that an ethical committee would have

approved our heart transplant study. Hearts are in short supply and society favors assigning them to individuals who are more likely to benefit from the transplant, rather than assigning them randomly among potential recipients. Also one could question the feasibility of the study even if ethical issues were ignored: double-blind assignment is impossible, individuals assigned to medical treatment may not resign themselves to forego a transplant, and there may not be compatible hearts for those assigned to transplant. Even if the study were feasible, it would still take several years to complete it, and decisions must be made in the interim. Frequently, conducting an observational study is the least bad option.

Technical Point 2.3

Equivalence of IP weighting and standardization. Assume that A is discrete with finite number of values and that $f[a|l]$ is positive for all l such that $\Pr[L=l]$ is nonzero. This *positivity* condition is guaranteed to hold in conditionally randomized experiments. Under positivity, the standardized mean for treatment level a is defined as $\sum_l \mathrm{E}[Y|A=a,L=l]\,\Pr[L=l]$ and the IP weighted mean of Y for treatment level a is defined as $\mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]$. The indicator function $I(A=a)$ is the function that takes value 1 for individuals with $A=a$, and 0 for the others.

We now prove the equality of the IP weighted and standardized means under positivity. By definition of expectation, $\mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right] = \sum_l \frac{1}{f[a|l]}\{\mathrm{E}[Y|A=a,L=l]\ f[a|l]\Pr[L=l]\} = \sum_l \{\mathrm{E}[Y|A=a,L=l]\Pr[L=l]\}$ where in the final step we cancelled $f[a|l]$ from the numerator and denominator, and in the first step we did not need to sum over the possible values of A because for any a' other than a the quantity $I(a')$ is zero. The proof treats A and L as discrete but not necessarily dichotomous. For continuous L simply replace the sum over L with an integral.

The proof makes no reference to counterfactuals. However, if we further assume conditional exchangeability, then both the IP weighted and the standardized means are equal to the counterfactual mean $\mathrm{E}[Y^a]$. Here we provide two different proofs of this last statement. First, we prove equality of $\mathrm{E}[Y^a]$ and the standardized mean as in the text: $\mathrm{E}[Y^a] = \sum_l \mathrm{E}[Y^a|L=l]\Pr[L=l] = \sum_l \mathrm{E}[Y^a|A=a,L=l]\Pr[L=l] = \sum_l \mathrm{E}[Y|A=a,L=l]\Pr[L=l]$ where the second equality is by conditional exchangeability and positivity, and the third by consistency. Second, we prove equality of $\mathrm{E}[Y^a]$ and the IP weighted mean as follows: $\mathrm{E}\left[\frac{I(A=a)}{f[A|L]}Y\right]$ is equal to $\mathrm{E}\left[\frac{I(A=a)}{f[A|L]}Y^a\right]$ by consistency. Next, because positivity implies $f[a|L]$ is never 0, we have

$\mathrm{E}\left[\frac{I(A=a)}{f[A|L]}Y^a\right] = \mathrm{E}\left\{\mathrm{E}\left[\frac{I(A=a)}{f[a|L]}Y^a\middle|L\right]\right\} = \mathrm{E}\left\{\mathrm{E}\left[\frac{I(A=a)}{f[a|L]}\middle|L\right]\mathrm{E}[Y^a|L]\right\}$ (by conditional exchangeability)

$= \mathrm{E}\{\mathrm{E}[Y^a|L]\}$ (because $\mathrm{E}\left[\frac{I(A=a)}{f[a|L]}\middle|L\right] = 1$) $= \mathrm{E}[Y^a]$.

When treatment is continuous, which is an unlikely design choice in conditionally randomized experiments, $\mathrm{E}[I(A=a)Y/f(A|L)]$ is no longer equal to $\sum_l \mathrm{E}[Y|A=a,L=l]\Pr[L=l]$ and thus is biased for $\mathrm{E}[Y^a]$ even under exchangeability. To see this, one can calculate that $\mathrm{E}[I(A=a)/f(a|l)\,|L=l]$ is equal to 0 rather than 1 if we take $f(a|l)$ to be (a version of) the conditional density of A given $L=l$ (with respect to Lebesgue measure). On the other hand, if we continue to take $f(a|l)$ to be $\Pr[A=a|L=l]$, the denominator $f(a|L=l)$ is zero on a set with probability 1 so positivity fails. In Section 12.4 we discuss how IP weighting can be generalized to accomodate continuous treatments. In Technical Point 3.1, we discuss that the results above do not hold in the absence of positivity, even for discrete A.