3.1 Potential Outcomes

Let's begin by thinking about the philosophical concept of a *potential outcome*. Prior to some "cause" occurring, for example receiving some exposure, the *potential outcomes* are all of the potential things that could occur depending on what you are exposed to. For simplicity, let's assume an exposure has two levels:

- $X = 1$ if you are exposed
- $X = 0$ if you are not exposed

Under this simple scenario, there are two potential outcomes:

- $Y(1)$ the potential outcome if you are exposed
- $Y(0)$ the potential outcome if you are not exposed

Only *one* of these potential outcomes will actually be realized, the one corresponding to the exposure that actually occurred, and therefore only one is observable. It is important to remember that these exposures are defined at a particular instance in time, so only one can happen to any individual. In the case of a binary exposure, this leaves one potential outcome as *observable* and one *missing*. In fact, early causal inference methods were often framed as missing data problems; we need to make certain assumptions about the *missing counterfactuals*, the value of the potential outcome corresponding to the exposure(s) that did not occur.

Our causal effect of interest is often some difference in potential outcomes $Y(1) - Y(0)$, averaged over a particular population.

3.2 Counterfactuals

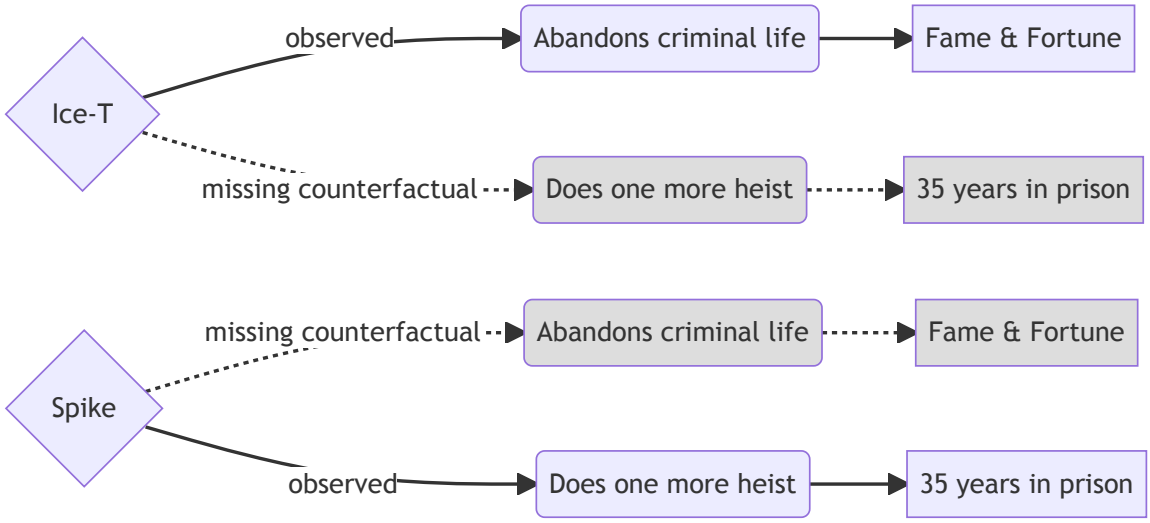
Conceptually, the missing counterfactual outcome is one that would have occurred under a different set of circumstances. In causal inference, we *wish* we could observe the conterfactual outcome that would have occurred in an alternate universe where the exposure status for a given observation was flipped. To do this, we attempt to control for all factors that are related to an exposure and outcome such that we can *construct* (or estimate) such a counterfactual outcome.

Let's think about a specific example. Ice-T, best known as an American rapper and Fin on Law and Order: SVU, co-authored a book titled "Split Decision: Life Stories", published in 2022. Here is the synopsis:

> **Award-winning actor, rapper, and producer Ice-T unveils a compelling memoir of his early life robbing jewelry stores until he found fame and fortune—while a handful of bad choices sent his former crime partner down an incredibly different path.**
>
> Ice-T rose to fame in the late 1980s, earning acclaim for his music before going on to enthrall television audiences as Odafin "Fin" Tutuola in *Law & Order: Special Victims Unit*. But it could have gone much differently.
>
> In this "poignant and powerful" (*Library Journal*, starred review) memoir, Ice-T and Spike, his former crime partner—collaborating with *New York Times* bestselling author Douglas Century—relate the shocking stories of their shared pasts, and how just a handful of decisions led to their incredibly different lives. Both grew up in violent, gang-controlled Los Angeles neighborhoods and worked together to orchestrate a series of jewelry heists.
>
> But while Ice-T was discovered rapping in a club and got his first record deal, Spike was caught for a jewelry robbery and did three years in prison. As his music career began to take off, Ice made the decision to abandon the criminal life; Spike continued to plan increasingly ingenious and risky jewel heists. And in 1992, after one of

> Spike's robberies ended tragically, he was sentenced to thirty-five years to life. While he sat behind bars, he watched his former partner rise to fame in music, movies, and television.
>
> "Propulsive" (*Publishers Weekly*, starred review), timely, and thoughtful, two men with two very different lives reveal how their paths might have very well been reversed if they made different choices. All it took was a *split decision*. ("Split Decision" 2022)

This premise is compelling because it implies that we are observing a *counterfactual*. The book begins by setting up all the ways Ice-T and his friend Spike were similar prior to some important moment (both grew up in Los Angeles neighborhoods, both were involved with gangs, both worked together to orchestrate a series of jewelry heists, etc). Then something happens – Ice-T makes a decision to abandon criminal life and Spike makes the opposite decision. What happens next for Ice-T includes fame and fortune, while Spike ends up with 35 years to life in prison. This book is attempting a small study, two people who prior to some event were the same and after were different – Spike's outcomes serve as the counterfactual to Ice-T's.



Ice-T and Spike Causal Map

In practice, this is what we attempt to do with causal inference techniques. Even randomized trials are limited to a single factual world, so we compare the average effects of groups with different exposures. Now, having this as a concrete example of an attempt to construct a counterfactual scenario in the "real-world" there are several issues that we can immediately see, highlighting the difficulty in drawing such inference. First, while the synopsis implies that the two individuals were similar prior to the precipitating event that dictated their future opposite directions, we can easily identify factors in which perhaps they differed. Ice-T decided to leave his life of crime, but that wasn't the only factor in his success: he had enough musical talent to make a career of it. Did Spike have Ice-T's musical talent? Can we really conclude that his life would have turned out exactly like Ice-T's if he had made the exact same choices as Ice-T? If we want to truly estimate the causal effect of the decision to leave criminal life on Ice-T's future outcomes, we would need to observe his ultimate course both under making the decision and not. Of course this is not possible, so what can we do? Perhaps we can find someone else who is exactly like Ice-T who did not make the same decision and see how they fare. Of course, Ice-T is unique, it would be challenging to find someone exactly like him. Again, this is attempted with Spike, and even so presents challenges. Often, instead of relying on a single individual, we rely on many individuals. We could conduct an experiment where we *randomize* many individuals to leave criminal life (or not) and see how this impacts their outcomes *on average* (this randomized trial seems to present some ethical issues, perhaps we need to look to *observational* studies to help answer this question). In any case, we must rely on statistical techniques to help construct these unobservable counterfactuals.

3.2.1 Potential Outcomes Simulation

Let's suppose some happiness index, from 1-10 exists. We are interested in assessing whether eating chocolate ice cream versus vanilla will increase happiness. We have 10 individuals with two potential outcomes for each, one is what

their happiness would be if they ate chocolate ice cream, (defined as `y_chocolate` in the code below), and one is what their happiness would be if they ate vanilla ice cream (defined as `y_vanilla` in the code below). We can define the true causal effect of eating chocolate ice cream (versus vanilla) on happiness for each individual as the difference between the two (Table 3.1).

```
data <- data.frame(
  id = 1:10,
  y_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
)

data <- data |>
  mutate(causal_effect = y_chocolate - y_vanilla)

data
```

	Potential Outcomes		Causal Effect
id	$Y_i(\text{chocolate})$	$Y_i(\text{vanilla})$	$Y_i(\text{chocolate}) - Y_i(\text{vanilla})$
1	4	1	3
2	4	3	1
3	6	4	2
4	5	5	0
5	6	5	1
6	5	6	-1
7	6	8	-2
8	7	6	1
9	5	3	2
10	6	5	1

Table 3.1: Potential Outcomes Simulation: The causal effect of eating chocolate (versus vanilla) ice cream on happiness

```
data |>
  summarize(
    avg_chocolate = mean(y_chocolate),
    avg_vanilla = mean(y_vanilla),
    avg_causal_effect = mean(causal_effect)
  )
```

```
  avg_chocolate avg_vanilla avg_causal_effect
1           5.4         4.6               0.8
```

For example, examining Table 3.1, the causal effect of eating chocolate ice cream (versus vanilla) for individual `4` is 0, whereas the causal effect for individual `9` is 2. The *average* potential happiness after eating chocolate is 5.4 and the *average* potential happiness after eating vanilla is 4.6. The *average* treatment effect of eating chocolate (versus vanilla) ice cream among the ten individuals in this study is 0.8.

In reality, we cannot observe both potential outcomes, in any moment in time, each individual in our study can only eat *one* flavor of ice cream. Suppose we let our participants choose which ice cream they wanted to eat and each choose their favorite (i.e. they knew which would make them "happier" and picked that one. Now what we *observe* is shown in Table 3.2.

```
data_observed <- data |>
  mutate(
```

```
    exposure = case_when(
      # people who like chocolate more chose that
      y_chocolate > y_vanilla ~ "chocolate",
      # people who like vanilla more chose that
      y_vanilla >= y_chocolate ~ "vanilla"
    ),
    observed_outcome = case_when(
      exposure == "chocolate" ~ y_chocolate,
      exposure == "vanilla" ~ y_vanilla
    )
  ) |>
  # we can only observe the exposure and one potential outcome
  select(id, exposure, observed_outcome)
data_observed
```

	Exposure	Observed Outcome
id	X_i	Y_i
1	chocolate	4
2	chocolate	4
3	chocolate	6
4	vanilla	5
5	chocolate	6
6	vanilla	6
7	vanilla	8
8	chocolate	7
9	chocolate	5
10	chocolate	6

Table 3.2: Potential Outcomes Simulation: The observed exposure and outcome used to estimate the effect of eating chocolate (versus vanilla) ice cream on happiness

```
data_observed |>
  group_by(exposure) |>
  summarise(avg_outcome = mean(observed_outcome))
```

```
# A tibble: 2 × 2
  exposure  avg_outcome
  <chr>           <dbl>
1 chocolate        5.43
2 vanilla          6.33
```

Now, the *observed* average outcome among those who ate chocolate ice cream is 5.4 (the same as the true average potential outcome), while the *observed* average outcome among those who ate vanilla is 6.3 – quite different from the *actual* average (4.6). The estimated causal effect here could be calculated as 5.4 - 6.3 = -0.9.

It turns out here, these 10 participants *chose* which ice cream they wanted to eat and they always chose to eat their favorite! This artificially made it look like eating vanilla ice cream would increase the happiness in this population when in fact we know the opposite is true. The next section will discuss which assumptions need to be true in order to allow us to *accurately* estimate causal effects using observed data. As a sneak peak, our issue here was that how the exposure was decided, if instead we *randomized* who ate chocolate versus vanilla ice cream we would (on average, with a large enough sample) recover the true causal effect.

```r
## we are doing something *random* so let's set a seed so we always observe the
## same result each time we run the code
set.seed(11)
data_observed <- data |>
  mutate(
    # change the exposure to randomized, generate from a binomial distribution
    # with a probability 0.5 for being in either group
    exposure = case_when(
      rbinom(10, 1, 0.5) == 1 ~ "chocolate",
      TRUE ~ "vanilla"
    ),
    observed_outcome = case_when(
      exposure == "chocolate" ~ y_chocolate,
      exposure == "vanilla" ~ y_vanilla
    )
  ) |>
  # we can only observe the exposure and one potential outcome
  select(id, exposure, observed_outcome)
data_observed |>
  group_by(exposure) |>
  summarise(avg_outcome = mean(observed_outcome))
```

```
# A tibble: 2 × 2
  exposure  avg_outcome
  <chr>           <dbl>
1 chocolate        5.33
2 vanilla          4.71
```

3.3 Causal Assumptions

Like most statistical approaches, the validity of a causal analysis depends on how well certain assumptions are met. As mentioned in Section 3.1, the potential outcomes framework envisions that each individual possesses a range of potential outcomes for every conceivable value of some input. For instance, as in the scenario previously described with two exposure levels (exposed: 1 and unexposed: 0), we can define potential outcomes for exposure ($Y(1)$) and no exposure ($Y(0)$), and subsequently analyze the difference between these outcomes, i.e., $Y(1) - Y(0)$, to comprehend the impact of the input (the exposure) on the outcome, Y. At any given time, only one of these *potential outcomes* is observable – namely, the outcome tied to the actual exposure the individual underwent. Under certain assumptions, we can leverage data from individuals exposed to different inputs to compare the average differences in their observed outcomes. The most common assumptions across the approaches we describe in this book are:

1. **Consistency**: We assume that the causal question you claim you are answering is consistent with the one you are *actually* answering with your analysis. Mathematically, this means that $Y_{obs} = (X)Y(1) + (1 - X)Y(0)$, in other words, the outcome you observe is exactly equal to the potential outcome under the exposure you received. Two common ways to discuss this assumption are:
    - **Well defined exposure**: We assume that for each value of the exposure, there is no difference between subjects in the delivery of that exposure. Put another way, multiple versions of the treatment do not exist.
    - **No interference**: We assume that the outcome (technically all *potential* outcomes, regardless of whether they are observed) for any subject does not depend on another subject's exposure.

> **Jargon**
>
> Assumption 1 is sometimes referred to as *stable-unit-treatment-value-assumption* or SUTVA (Imbens and Rubin 2015). Likewise, these assumptions are sometimes referred to as *identifiability conditions* since we need them to hold in order to identify causal estimates.

2. **Exchangeability**: We assume that within levels of relevant variables (confounders), exposed and unexposed subjects have an equal likelihood of experiencing any outcome prior to exposure; i.e. the exposed and unexposed subjects are exchangeable. This assumption is sometimes referred to as **no unmeasured confounding**.

3. **Positivity**: We assume that within each level and combination of the study variables used to achieve exchangeability, there are exposed and unexposed subjects. Said differently, each individual has some chance of experiencing every available exposure level. Sometimes this is referred to as the **probabilistic** assumption.

> **Apples-to-apples**
>
> Practically, most of the assumptions we need to make for causal inference are so we can make an *apples-to-apples* comparison: we want to make sure we're comparing individuals that are similar — who would serve as good proxies for each other's counterfactuals.
>
> The phrase *apples-to-apples* stems from the saying "comparing apples to oranges", e.g. comparing two things that are incomparable.
>
> That's only one way to say it. There are a lot of variations worldwide. Here are some other things people incorrectly compare:
>
> - Cheese and chalk (UK English)
> - Apples and pears (German)
> - Potatoes and sweet potatoes (Latin American Spanish)
> - Grandmothers and toads (Serbian)
> - Horses and donkeys (Hindi)

3.3.1 Causal Assumptions Simulation

Let's bring back our simulation from Section 3.2.1. Recall that we have individuals who will either eat chocolate or vanilla ice cream and we are interested in assessing the causal effect of this exposure on their happiness. Let's see how violations of each assumption may impact the estimation of the causal effect.

3.3.1.1 Consistency violation

Two ways the consistency assumption can be violated is (1) lack of a well defined exposure and (2) interference. Let's see how these impact our ability to accurately estimate a causal effect.

3.3.1.1.1 Well defined exposure

Suppose that there were in fact two containers of chocolate ice cream, one of which was spoiled. Therefore, despite the fact that having an exposure "chocolate" could mean different things depending on where the individual's scoop came from (regular chocolate ice cream, or spoiled chocolate ice cream), we are lumping them all together under a single umbrella (hence the violation, we have "multiple versions of treatment"). You can see how this falls under consistency because the issue here is that the potential outcome we think we are estimating is not the one we are actually observing.

```
data <- data.frame(
  id = 1:10,
  y_spoiledchocolate = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
  y_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
) |>
  mutate(causal_effect = y_chocolate - y_vanilla)

set.seed(11)
data_observed <- data |>
  mutate(
    exposure_unobserved = case_when(
      rbinom(10, 1, 0.25) == 1 ~ "chocolate (spoiled)",
      rbinom(10, 1, 0.25) == 1 ~ "chocolate",
      TRUE ~ "vanilla"
    ),
    observed_outcome = case_when(
      exposure_unobserved == "chocolate (spoiled)" ~ y_spoiledchocolate,
      exposure_unobserved == "chocolate" ~ y_chocolate,
      exposure_unobserved == "vanilla" ~ y_vanilla
    ),
```

```
    exposure = case_when(
      exposure_unobserved %in% c("chocolate (spoiled)", "chocolate") ~ "chocolate",
      exposure_unobserved == "vanilla" ~ "vanilla"
    )
  ) |>
  select(id, exposure, observed_outcome)

data_observed |>
  group_by(exposure) |>
  summarise(avg_outcome = mean(observed_outcome))
```

```
# A tibble: 2 × 2
  exposure  avg_outcome
  <chr>           <dbl>
1 chocolate        2.75
2 vanilla          4.67
```

We know the *true* average causal effect of (unspoiled) chocolate in the sample is 0.8, however our estimated causal effect (because our data are not consistent with the question we are asking) is -1.9. This demonstrates what can go wrong when *well defined exposure* is violated.

3.3.1.1.2 Interference

Interference would mean that an individual's exposure impacts another's potential outcome. For example, let's say each individual has a partner, and their potential outcome depends on both what flavor of ice cream they ate *and* what flavor their partner ate. For example, in the simulation below, having a partner that received a different flavor of ice cream increases the happiness by two units.

```
data <- data.frame(
  id = 1:10,
  partner_id = c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5),
  y_chocolate_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_chocolate_vanilla = c(6, 6, 8, 7, 8, 7, 8, 9, 7, 8),
  y_vanilla_chocolate = c(3, 5, 6, 7, 7, 8, 10, 8, 5, 7),
  y_vanilla_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
)

set.seed(11)
data_observed <- data |>
  mutate(
    exposure = case_when(
      rbinom(10, 1, 0.5) == 1 ~ "chocolate",
      TRUE ~ "vanilla"
    ),
    exposure_partner =
      c("vanilla", "vanilla", "vanilla", "chocolate", "chocolate", "vanilla", "vanilla", "vanilla", "va
    observed_outcome = case_when(
      exposure == "chocolate" & exposure_partner == "chocolate" ~ y_chocolate_chocolate,
      exposure == "chocolate" & exposure_partner == "vanilla" ~ y_chocolate_vanilla,
      exposure == "vanilla" & exposure_partner == "chocolate" ~ y_vanilla_chocolate,
      exposure == "vanilla" & exposure_partner == "vanilla" ~ y_vanilla_vanilla
    )
  ) |>
  select(id, exposure, observed_outcome)

data_observed |>
  group_by(exposure) |>
  summarise(avg_outcome = mean(observed_outcome))
```

```
# A tibble: 2 × 2
  exposure  avg_outcome
  <chr>           <dbl>
1 chocolate        7.33
2 vanilla          5.57
```

Now our estimated causal effect (because interference exists) is 1.8. This demonstrates what can go wrong when *interference* occurs. One of the main ways to combat interference is change the *unit* under consideration. Here, each individual, each unique *id*, is considered a unit, and there is interference between units (i.e. between partners). If instead we consider each *partner* as a unit and randomize the partners rather than the individuals, we solve the interference issue, as there is not interference *between* different partner sets. This is sometimes referred to as a *cluster randomized trial*. What we decide to do within each cluster may depend on the causal question at hand. For example, if we want to know what would happen if *everyone* at chocolate ice cream versus if *everyone* at vanilla, we would want to randomize both partners to either chocolate or vanilla, as seen below.

```
set.seed(11)

## we are now randomizing the *partners* not the individuals
partners <- data.frame(
  partner_id = 1:5,
  exposure = case_when(
    rbinom(5, 1, 0.5) == 1 ~ "chocolate",
    TRUE ~ "vanilla"
  )
)
data_observed <- data |>
  left_join(partners, by = "partner_id") |>
  mutate(
    # all partners have the same exposure
    exposure_partner = exposure,
    observed_outcome = case_when(
      exposure == "chocolate" & exposure_partner == "chocolate" ~ y_chocolate_chocolate,
      exposure == "vanilla" & exposure_partner == "vanilla" ~ y_vanilla_vanilla
    )
  ) |>
  select(id, exposure, observed_outcome)

data_observed |>
  group_by(exposure) |>
  summarise(avg_outcome = mean(observed_outcome))
```

```
# A tibble: 2 × 2
  exposure  avg_outcome
  <chr>           <dbl>
1 chocolate        5.5
2 vanilla          4.38
```

3.3.1.2 Exchangeability violation

We have actually already seen an example of an exchangeability violation in Section 3.2.1. In that example, participants were able to choose the ice cream that they wanted to eat, so people who were more likely to have a positive effect from eating chocolate chose that, and those more likely to have a positive effect from eating vanilla chose that.

```
data <- data.frame(
  id = 1:10,
  y_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
)
data_observed <- data |>
  mutate(
    exposure = case_when(
      # people who like chocolate more chose that
      y_chocolate > y_vanilla ~ "chocolate",
      # people who like vanilla more chose that
      y_vanilla >= y_chocolate ~ "vanilla"
    ),
    observed_outcome = case_when(
```

```
      exposure == "chocolate" ~ y_chocolate,
      exposure == "vanilla" ~ y_vanilla
    )
  ) |>
  select(id, exposure, observed_outcome)

data_observed |>
  group_by(exposure) |>
  summarise(avg_outcome = mean(observed_outcome))
```

```
# A tibble: 2 × 2
  exposure  avg_outcome
  <chr>           <dbl>
1 chocolate        5.43
2 vanilla          6.33
```

How could we correct this? If we had some people who preferred chocolate ice cream but ended up taking vanilla instead, we could *adjust* for the preference, and the effect conditioned on this would no longer have an exchangeability issue. It turns out that this example as we have constructed it doesn't lend itself to this solution because participants chose their preferred flavor 100% of the time making this *also* a positivity violation.

3.3.1.3 Positivity violation

As stated above, the previous example violates both *exchangeability* and *positivity*. How could we fix it? As long as *some* people chose outside their preference with some probability (even if it is small!) we can remove this violation. Let's say instead of everyone picking their flavor of preference 100% of the time, they just had a 80% chance of picking that flavor.

```
data <- data.frame(
  id = 1:10,
  y_chocolate = c(4, 4, 6, 5, 6, 5, 6, 7, 5, 6),
  y_vanilla = c(1, 3, 4, 5, 5, 6, 8, 6, 3, 5)
)

set.seed(11)
data_observed <- data |>
  mutate(
    prefer_chocolate = y_chocolate > y_vanilla,
    exposure = case_when(
      # people who like chocolate more chose that 80% of the time
      prefer_chocolate ~ ifelse(rbinom(10, 1, 0.8), "chocolate", "vanilla"),
      # people who like vanilla more chose that 80% of the time
      !prefer_chocolate ~ ifelse(rbinom(10, 1, 0.8), "vanilla", "chocolate")
    ),
    observed_outcome = case_when(
      exposure == "chocolate" ~ y_chocolate,
      exposure == "vanilla" ~ y_vanilla
    )
  ) |>
  select(id, prefer_chocolate, exposure, observed_outcome)

lm(
  observed_outcome ~ I(exposure == "chocolate") + prefer_chocolate,
  data_observed
)
```

```
Call:
lm(formula = observed_outcome ~ I(exposure == "chocolate") +
    prefer_chocolate, data = data_observed)

Coefficients:
                  (Intercept)
                        6.156
```

```
I(exposure == "chocolate")TRUE
                         0.531
          prefer_chocolateTRUE
                        -1.469
```

After *adjusting* for this variable (chocolate preference), we recover the correct causal effect. This value is not exactly the same as the truth we obtain with the (unobservable) potential outcomes because we are dealing with a small sample – as our sample size increases this will get closer to the truth.

Causal assumptions can be difficult to verify and may not hold for many data collection strategies. We cannot overstate the importance of checking these criteria to the extent possible! Following any of the recipes in this book are unlikely to give valid answers if the causal assumptions are badly violated.