Causal Inference

Ashley I Naimi

Spring 2022

Contents

1 Correlation and Causation

In the *The Grammar of Science,* Karl Pearson (1911) wrote: "[b]eyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect." He suggested that rather than pursue an understanding of cause-effect relations, scientists would be best served by measuring correlations through tables that classify individuals into specific categories. "Such a table is termed a contingency table, and the ultimate scientific statement of description of the relation between two things can always be thrown back upon such a contingency table."

Over a century later, a majority of statistics courses tend to treat causal inference by simply stating that "correlation is not causation." This treatment is hardly sufficient, for at least two reasons: 1) As scientists, our primary interest is (should be) in cause-effect relations; 2) People continue to conflate correlation with causation[1]. For both of these reasons, we very much need to **clarify the conditions that would allow us to understand causality better.** This is what "causal inference" is all about.

Generally, I adopt the view that **the causal and statistical aspects of a scientific study should be kept as separate as possible.** The objective is to first define the effect and articulate the conditions under which causal inference is possible for this effect, and then to understand what statistical tools will enable us to answer the causal question.[2] Causal inference tells us what we should estimate, and whether we can. Statistics tells us how to estimate it. By implication, we should avoid treating statistical models as if they were causal. For example, the practice of reading the risk ratio, odds ratio, or risk difference for an exposure of interest from a generalized linear (statistical) model[3] will sometimes work under very specific conditions, but is not the best approach for quantifying exposure effects (Naimi and Whitcomb, 2020).

[1] Daniel Westreich and I reviewed a book whose authors were so caught up in the allure of "Big Data", they thoroughly forgot that correlation $\neq$ causation. See Naimi and Westreich (2014)

[2] Loosely speaking: Causal inference is the "what?" Statistics is the "how?"

[3] or the hazard ratio from a Cox model, or the mean ratio from a Poisson model, or host of other types of regression models

2 Introduction to Causal Inference

"Causal inference" deals primarily with the **formal mechanisms by which we can combine data, assumptions, and models to interpret a correlation (or**

association) causally.[4] The framework in which we define what we mean by "causal relation" or "causal effect" is the **potential outcomes framework**.

A central notion in the potential outcomes framework is the counterfactual. This notion stems from the intuitive and informal practice of interpreting cause-effect relations as **circumstances (e.g., health outcomes) that would have arisen had things (e.g., exposures) been different**.

While this intuition serves an important purpose, it is not sufficient for doing rigorous science. Suppose we ask: "what is the effect of smoking on the 5-year cumulative CVD risk, irrespective of smoking's effect on body weight?" This question may seem clear and intuitive. To answer this question, we would do a study in which we collect data, enter these into a computer, perform some calculations, and obtain a number (we'd usually like to interpret as the "effect").

But there is a problem.[5] The calculations performed by the computer are **rigorously defined (i.e., unambiguous) mathematical objects**. On the other hand, **English language sentences about cause effect relations are ambiguous**. For example, the "effect of smoking" can mean many different things:

- All people smoke any tobacco ever versus no people smoke tobacco ever.
- All people smoke 3 cigarettes per day versus all people smoke 2 cigarettes per day.
- All people who have smoked any tobacco in the last 15 years cease to smoke any tobacco whatsoever.

Similarly, "irrespective of" can mean a number of things:

- The effect of smoking on CVD risk that would be observed in a hypothetical world where smoking did not affect body mass?
- The effect of smoking on CVD risk if everyone were set to "normal" body mass?
- The effect of smoking on CVD risk if everyone were held at the body mass they had in the month prior to study entry?

But the numerical strings of data and the computer algorithms applied to these data are well defined mathematical objects, which do not admit such ambiguity. Depending on several choices, including the data, how variables are coded, and the modeling strategy, the computer is being told which question to answer. There is a lot of potential uncertainty in the space between the English

[4] There are a number of introductory books and articles on causal inference in the empirical sciences. Here are some excellent options: Hernán and Robins (2020), Pearl et al. (2016), Imbens and Rubin (2015), Cunningham (2021)

[5] This problem was articulated by Robins 1987, and I am using a version of the example from his paper.

language sentences we use to ask causal questions, and the computer algorithms we use to answer those questions. Causal inference is about clarifying this uncertainty.

3 Potential Outcomes Notation

The building blocks for causal inference are **potential outcomes** (Rubin, 2005).

Importantly, these are conceptually distinct from **observed outcomes**. That is, the outcome that one might observe in a dataset is not the same as the potential outcome.

Potential outcomes are functions of exposures. For a given exposure x, we will write the potential outcome as Y^x.[6] **This is interpreted as "the outcome (Y) that would be observed if X were set to some value x"**. For example, if X is binary [denoted $X \in (0, 1)$], then Y^x is the outcome that would be observed if $X = 0$ or $X = 1$. If we wanted to be specific about the value of x, we could write $Y^{x=0}$ or $Y^{x=1}$ (or, more succinctly, Y^0 or Y^1).

[6] Alternate notation includes: Y_x, $Y(x)$, $Y \mid Set(X = x)$, and $Y|do(X = x)$.

Concept Question:

Suppose you collect data from a single person and find that they are exposed. Can you interpret the outcome you observe to be the potential outcome that would have been observed had they been exposed? Why or why not?

4 Estimand, Estimator, Estimate

4.1 Estimand

Causal inference starts with a clear idea of the effect of interest (the target causal parameter, or **estimand**). We use potential outcomes to do this, but it is useful and important first to distinguish between estimands, estimators, and estimates. The **estimand** is the (mathematical) object we want to quantify. It is, for example, the causal risk difference, risk ratio, or odds ratio for our exposure and outcome of interest. In our smoking CVD example, we might be interested in:

$$E(Y^1 - Y^0), \quad \frac{E(Y^1)}{E(Y^0)}, \quad \frac{Odds(Y^1 = 1)}{Odds(Y^0 = 1)},$$

where $Odds(Y^x = 1) = E(Y^x)/[1 - E(Y^x)]$, and where $E(.)$ is the expectation operator taken with respect to the total population.[7] There are many other causal estimands besides these (effect of treatment on the treated, complier average causal effect, survivor average causal effect, stochastic effects, other).

Furthermore, the estimand need not always be causal (Casella and Berger, 2002). We may be interested in a statistical estimand, such as the conditional risk difference, risk ratio, or odds ratio:

$$E(Y \mid X = 1) - E(Y \mid X = 0), \quad \frac{E(Y \mid X = 1)}{E(Y \mid X = 0)}, \quad \frac{Odds(Y \mid X = 1)}{Odds(Y \mid X = 0)},$$

What's important is that one is clear about the objective. For example, in Naimi (2016) we defined counterfacutal disparity measures as:

$$E(Y^m \mid X = 1) - E(Y^m \mid X = 0)$$

which is a mixed statistical and counterfactual estimand. It is a measure of disparity (statistical estimand) that would be observed if some variable M were set to a value m (counterfactual estimand).

The causal estimands presented above represent **average treatment effects** (on the risk difference, risk ratio, and odds ratio scale, respectively). This effect is sometimes referred to as a marginal treatment effect, because it averages (or marginalizes) the effect over the entire sample. For instance, if we consider the risk difference, it is easy to show that[8]

$$E(Y^1 - Y^0) = \frac{1}{N}\sum_{i=1}^{N} Y_i^1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0$$

However, we may want to estimate this effect in a subset of the population. For instance, $E(Y^1 - Y^0 \mid C = c)$ is the effect of $x = 1$ versus $x = 0$ among those with $C = c$. There are many different conditional treatment (in contrast to marginal) effects, this latter one being one of the simplest. Another common conditional treatment effect is the effect of treatment on the treated (ETT):

$$E(Y^1 - Y^0 \mid X = 1)$$

This effect compares the outcomes that would be observed if the exposure

[7] Throughout this course, if the outcome Y is binary, then $E(Y) \equiv P(Y = 1)$. Or, the expectation of Y is equivalent to the probability that $Y = 1$. This assumes that the binary outcome variable Y is coded as $\{0, 1\}$, and not, e.g., $\{1, 2\}$. For the more technically oriented,

$$E(Y) = \int y f(y) dy$$

where $f(y)$ is the probability density function of Y.

[8] Recall that Y^x is not the observed (or sample) value of the outcome, so how do we actually get this average? When we discuss identifiability, we will see how we use observed data to quantify these contrasts.

were set to 1 (Y^1) versus if the exposure were set to 0 (Y^0) among those who were observed to be or actually exposed in the sample ($X = 1$).

To illustrate the relevance of this effect, consider the following (entirely fictional) scenario: Suppose that during gestation of a high-risk pregnancy, two clinical options are available to manage the risk of fetal death: premature delivery induction versus expectant management. Suppose further a researcher is interested in quantifying the effect of inducing delivery prematurely on fetal and infant death. This researcher collects data on a cohort of high-risk pregnant women, including whether delivery was induced prematurely, fetal/infant death, and a host of confounding variables. All parties involved agree the study is designed perfectly (no confounding, measurement error, loss to follow-up). They calculate the average treatment effect of premature delivery induction on fetal and infant death on the risk difference scale:

$$E(Y^1 - Y^0) = 0.15$$

This researcher concludes that, if all high-risk pregnancies were induced prematurely ($X = 1$), 15 more out of every 100 pregnancies would end in death, relative to what would happen if all high-risk pregnancies were left to expectant management ($X = 0$). In light of this incredibly high excess risk of death, this researcher advises abandoning the practice of premature delivery induction entirely.

Another researcher questions the relevance of the average treatment effect. They argue that physicians would never induce delivery prematurely in all versus no high-risk pregnancies. Rather, the more interesting question is: **for those women whose pregnancies were actually induced**, what would the risk of death have been had they not been induced? Underlying this more nuanced question is an understanding that physicians may be inducing pregnancy in women because of a number of reasons that make these women different from the rest. These differences, in turn, can lead to a different effect. This researcher thus calculates the effect of treatment on the treated:

$$E(Y^1 - Y^0 \mid X = 1) = -0.05$$

This other researcher concludes that, among those whose pregnancies were actually delivered prematurely, the risk of death would have been higher had

they not been delivered prematurely.

This hypothetical example demonstrates a fundamental difference between the ATE and the ETT: for those high-risk pregnancies that were not induced prematurely, the act of inducing premature delivery would not be beneficial. But for those high-risk pregnancies that were induced prematurely, the act of inducing premature delivery was beneficial. The ATE averages the beneficial and non-beneficial effects in the entire population, to yield an overall non-beneficial effect. The ETT isolates the beneficial effect among those who actually received the intervention. Thus, in this hypothetical example, premature delivery actually did benefit those who received it, even though it would not benefit everybody.

There are many other estimands that can be defined, including the local average treatment effect Angrist et al. (1996), the survivor average causal effect Tchetgen Tchetgen (2014), the complier average causal effect Shrier et al. (2014), principal strata effects Frangakis and Rubin (2002), stochastic effects Munoz and van der Laan (2012), incremental propensity score effects Naimi et al. (2021), and others. We will not discuss these in the context of this course, but it's good to be aware of their existence.

4.2 Estimator

The estimand is the object we want to estimate. The **estimator** is an equation that allows us to use our data to quantify the estimand. Suppose, for example, we were explicitly interested in quantifying the causal risk difference for the relation between smoking and 5 year CVD risk. To do this, we **have to** start by quantifying the associational risk difference, but there are many ways to do this (e.g., ordinary least squares, maximum likelihood, or many others).

To be specific, let's simulate some hypothetical data on the relation between smoking and CVD. Let's look at ordinary least squares, maximum likelihood, the generalized method of moments, and augmented inverse probability weighting (AIPW) as estimators:

```
remotes::install_github("yqzhong7/AIPW")
library(AIPW)
```

```
install.packages("SuperLearner",repos = "https://cloud.r-project.org/", dependencies=TRUE)
```

```
##
## The downloaded binary packages are in
##  /var/folders/z_/cty0tpg97wz_x1d1zgdhwllr0000gs/T//RtmpeD8yI7/downloaded_packages
```

```
library(SuperLearner)

# define the expit function
expit<-function(z){1/(1+exp(-(z)))}
set.seed(123)
n<-1e6
confounder<-rbinom(n,1,.5)
smoking<-rbinom(n,1,expit(-2+log(2)*confounder))
CVD<-rbinom(n,1,.1+.05*smoking+.05*confounder)

# the data
head(data.frame(CVD,smoking,confounder))
```

```
##   CVD smoking confounder
## 1   0       0          0
## 2   0       0          1
## 3   1       0          0
## 4   1       0          1
## 5   0       0          1
## 6   0       0          0
```

```
round(mean(confounder),3)
```

```
## [1] 0.499
```

```
round(mean(smoking),3)
```

```
## [1] 0.166
```

```
round(mean(CVD),3)
```

```
## [1] 0.133
```

```
#OLS
round(coef(lm(CVD~smoking+confounder)),4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0485      0.0501
```

```
#ML1
round(coef(glm(CVD~smoking+confounder,family=poisson("identity"))),4)
```

```
## (Intercept)     smoking  confounder
##      0.0999      0.0487      0.0502
```

```
#ML2
round(coef(glm(CVD~smoking+confounder,family=binomial("identity"))),4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0487      0.0501
```

```
#GMM
# round(gmm(CVD~smoking+confounder,x=cbind(smoking, confounder))$coefficients,4)

#AIPW
AIPW_SL <- AIPW$new(Y = CVD,
                    A = smoking,
                    W = confounder,
                    Q.SL.library = c("SL.mean","SL.glm"),
                    g.SL.library = c("SL.mean","SL.glm"),
                    k_split = 3,
                    verbose=FALSE)$
  fit()$
```

```
  summary()


round(AIPW_SL$result[3,1],4)
```

```
## [1] 0.0488
```

In our simple setting with 1 million observations, ordinary least squares, maximum likelihood, the generalized method of moments, and AIPW yield the same associational risk difference (as expected) even though they are (for some, completely) different **estimators**.[9]

It is important to note that these estimates are not causal risk differences, but are associational. Even the results from the AIPW estimator are *associational*, even though this method is much more clearly motivated from within the causal inference framework (Robins and Greenland, 1994). To interpret them as causal effects, we have to evaluate whether we can **identify** the effect we want to estimate. We discuss this next.

[9] A slightly deeper dive into these concepts can be found in Naimi and Whitcomb (2020) Estimating Risk Ratios and Risk Differences Using Regression. Am J Epidemiol. 189(6):508-10

4.3 Estimates

Finally, the values obtained from each estimation approach (~0.05) are our **estimates**.

5 Identifiability: Average Treatment Effect

In our simulation example, we estimated the associational (as opposed to causal) risk difference using four different estimators (ordinary least squares, two different maximum likelihood estimators, and AIPW). Estimating associations is all we can do with empirical data. Any time you use software to obtain a point estimate, you get an associational measure, irrespective of the method used.[10]

But our primary interest is often in causal quantities. In our simulated case, we want to estimate the causal risk difference for the effect of smoking on CVD. We can only do so if this causal risk difference is **identified**. Formally, *a parameter (e.g., causal risk difference) is identified if we can write it as a function of the observed data.*

[10] This is true with ANY estimator, including IP-weighting, g computation, g estimation, or double robust approaches, such as AIPW (as demonstrated) or targeted maximum likelihood estimation.

The causal risk difference is defined as a contrast of potential outcomes. Referring back to our simulated example,[11] we want to estimate the causal risk difference which is an example of an average treatment effect:

$$E(Y^1 - Y^0),$$

where Y^1, Y^0 are the potential CVD outcomes that would be observed if smoking were set to 1 and 0, respectively. On the other hand, the associational risk difference is defined as a contrast of observed outcomes:

$$E(Y \mid X = 1) - E(Y \mid X = 0),$$

where each term in this equation is interpreted as the risk of CVD **among those who had** $X = x$.

The causal risk difference is identified if the following equation holds:

$$E(Y^x) = E(Y \mid X = x).$$

This equation says that the risk of CVD that would be observed if everyone were set to $X = x$ is equal to the risk of CVD that we observe among those with $X = x$. In this equation, the right hand side equation is written entirely in terms of observed data ($Y = 1$). The left hand side is a function of unobserved potential outcomes ($Y^x = 1$). Because potential outcomes are unobservable abstractions, this equivalence will only hold if we can make some assumptions.

[11] To simplify the explanation here, I am ignoring the fact that we conditioned on (or adjusted for) confounders C. Of course, without adjusting for C, we get a confounded estimate. However, if we adjust for C, we no longer obtain the average treatment effect. Instead, we obtain the conditional treatment effect. Their are important distinctions between average and conditional treatment effects that we will discuss in a subsequent section.

5.1 Counterfactual Consistency

The first is **counterfactual consistency**, which states that the potential outcome that would be observed if we set the exposure to the observed value is the observed outcome (Hernán, 2005, Hernan and Taubman (2008),Hernán and VanderWeele (2011),VanderWeele and Hernán (2013)).[12] Formally, counterfactual consistency states that:

$$\text{if } X = x \text{ then } Y^x = Y$$

The status of this assumption remains unaffected by the choice of analytic method (e.g., standard regression versus g methods). Rather, this assumption's

[12] While somewhat convoluted, this assumption is primarily about legitimizing the connection between our observational study, and future interventions in actual populations based on this study. In our observational study, we **see** people with with a certain value of the exposure. In a future intervention, we **set** people to a certain value of the exposure. The differences between seeing and setting can be profound.

validity depends on the nature of the exposure assignment mechanism.

One way to grasp what counterfactual consistency is about is to use the example of the "effect" of obesity on mortality (Hernan and Taubman, 2008). We know that obesity is associated with an increased risk of mortality, but interpreting this excess risk into a causal statement is tricky. In an observational study, the association between obesity and mortality is obtained by contrasting the risk of mortality among, say, obese versus non-obese individuals. However, causally acting on this information would require us to find a way to make obese individuals non-obese. This might consist of getting obese individuals to diet, exercise, start smoking, or to undergo a single leg amputation (!). Each of these interventions could reduce BMI, and thus getting obese individuals to become non-obese. However, each intervention will likely have (dramatically) different effects on mortality.

They key here is that obesity is not a manipulable construct (on the other hand, dieting, exercise, smoking, and leg amuptation, more or less, are). As a result, precisely translating what we mean by "the effect of obesity" is difficult. The same problem arises with other variables, such as the "effect of education," the "effect of race/ethnicity," and the "effect of socieoeconomic status," to name a few (Naimi and Kaufman, 2015).

5.2 Interference

We must also assume **no interference**, which states that the potential outcome for any given individual does not depend on the exposure status of another individual (Hudgens and Halloran, 2008, Naimi and Kaufman (2015)). If this assumption were not true, we would have to write the potential outcomes as a function of the exposure status of multiple individuals. For example, for two different people indexed by i and j, we might write: $Y_i^{x_i,x_j}$.[13] Notation and methods that account for interference can become very complex very quickly (Tchetgen Tchetgen and VanderWeele, 2012, Halloran and Hudgens (2016), Hudgens and Halloran (2008)). As a result, we will not consider the impact of interference here, except only to say that different estimands and estimators should be used to properly account for them.

[13] Together, counterfactual consistency and no interference make up the stable-unit treatment value assumption (SUTVA), first articulated by Rubin (1980).

Together, counterfactual consistency and no interference allow us to make some progress in writing the potential risk $E(Y^x)$ as a function of the observed risk $E(Y \mid X = x)$. Specifically, by counterfactual consistency and no

interference, we can do the following:

$$E(Y^x) = E(Y \mid X = x) \tag{1}$$
$$= E(Y^x \mid X = x) \tag{2}$$

5.3 Exchangeability

A third assumption is **exchangeability**, which implies that the potential outcomes under a specific exposure (Y^x) are independent of the observed exposures X (Greenland and Robins, 1986, Greenland et al. (1999),Greenland and Robins (2009)). To explain the intuition behind exchangeability (Hernán and Robins, 2020), consider a setting in which we are estimating the effect of aspirin on headache incidence in a cohort of individuals aged 18-40 years.[14] To do this experiment, a researcher randomly assigns 50% of the cohort to aspirin, and the remaining 50% to placebo. However, to overcome some logistical complications, before actually giving them aspirin/placebo, this researcher hands out cards that indicate whether the participant was assigned to aspirin (red card) versus placebo (blue card).

[14] Assume that our sample size is sufficiently large so as to avoid any sampling variability problems.

After the cards/aspirin/placebo are distributed and the follow-up period transpires, the researcher tallies up the number of headaches in each exposure group. He finds the following results:

$$\text{Aspirin (Red Card): } E(Y \mid X = 1) = 0.6$$
$$\text{Placebo (Blue Card): } E(Y \mid X = 0) = 0.1$$

However, after reviewing the study protocol, he realizes that he accidentally assigned placebo to those with the red card, and aspirin to those with the blue card, instead of the other way around. Fortunately, this has no actual impact on the study, with the exception of needing to switch the aspirin label with the placebo label. Why? Randomization (in a sufficiently large enough sample) creates independencies between outcome that would be observed under some exposure value (the potential outcome) and the observed exposure. In our case, $E(Y^{x=1}) = 0.1$, and this is the case whether the exposure received was

placebo ($X = 0$) or aspirin ($X = 1$):

$$E(Y^{x=1}) = 0.1 \implies \begin{cases} E(Y^{x=1} \mid X = 1) = 0.1 \\ E(Y^{x=1} \mid X = 0) = 0.1 \end{cases}$$

Thus, because of randomization the following mathematical relation is implied:

$$E(Y^x \mid X) = E(Y^x) \tag{3}$$

which is exactly what we need to progress the identifiability statement above:

$$\begin{aligned} E(Y^x) &= E(Y \mid X = x) && (4) \\ &= E(Y^x \mid X = x) \text{ by consistency and no interference} && (5) \\ &= E(Y^x) \text{ by exchangeability} && (6) \end{aligned}$$

Study Question:
Why is the word "exchangeable" used to describe this concept? What, precisely, is being "exchanged"?

5.4 Conditional Exchangeability

With exchangeability, we are able to drop the observed exposure on the right side of the conditioning statement. However, we motivated this exchangeability assumption via simple randomization. What about when we have an observational study where the exposure is not randomized? It turns out that the validity of results from an observational study still rests upon the idea of randomization. For example, if we conduct an analysis in observational data where we adjust for 3 confounding variables, and we believe these three variables are sufficient to control for all confounding (and there are no other threats to validity, such as selection or information bias), then we can show that the same set of steps required to equate the average potential outcomes $E(Y^x)$ with the average observed outcome among those with $X = x$: $E(Y \mid X = x)$.

Consider our aspirin and headache example above, instead rather than randomly assign 50% of the individuals to aspirin and 50% to placebo, imagine

that for people who in an average week sleep < 7 hours per night, we use a coin that chooses heads 75% if the time to assign aspirin, and 25% of the time to assign placebo. And for people who sleep $\geq$ 7 hours per night, we use a 50:50 coin to assign aspirin and placebo.

Using an aspirin:placebo assignment proportion of 75:25 for "non-sleepers", and 50:50 for "sleepers" creates an association between sleeping quantity and aspirin assignment. If sleeping quantity also has an association with headache, what we've done is created a confounding relation between aspirin versus placebo and headache via sleeping quantity. Because of this confounding relation, we can no longer re-write the conditional expectation $E(Y^x \mid X = x)$ as $E(Y^x)$.

However, if we adjust for sleeping quantity in our analysis, we can partly recover the procedure we need to equate these quantities:

$$E(Y^x) = \sum_c E(Y \mid X = x, C) \tag{7}$$

$$= \sum_c E(Y^x \mid X = x, C) \text{ by consistency and no interference} \tag{8}$$

$$= \sum_c E(Y^x \mid C) \text{ by conditional exchangeability} \tag{9}$$

$$= E(Y^x) \text{ by marginalization} \tag{10}$$

The only difference is that now we have to incorporate an additional step in which we "average" or marginalize over the distribution of C to obtain a weighted average of the $E(Y^x)$ in the sample or population.

Technical Note:

Consider the marginalization step in the identification equation above. This step involves transitioning from $\sum_c E(Y^x \mid C)$ to $E(Y^x)$. This simply denotes taking a weighted sum of $E(Y^x \mid C)$, where the weights are defined as a probability function of C. For example, if $C \in \{0, 1, \ldots k\}$, then this sum becomes:

$$E(Y^x \mid C = 0)P(C = 0) + E(Y^x \mid C = 1)P(C = 1) + \ldots + E(Y^x \mid C = k)P(C = k)$$

More generally (i.e., for a more general case where C is not necessarily categorical), we can rewrite this as:

$$E(E(Y^x \mid C))$$

where the outer expectation is taken over C, and the inner expectation is taken over Y^x. This equation is sometimes referred to as the law of iterated expectations, the law of total expectation, or the tower rule. It plays an important role in causal inference, such as when we define (and sometimes implement) the g computation estimator. It is useful to understand, both when reading the technical literature, as well as when implementing variations of the technique in software.

5.5 Positivity

Although it seems that we have successfully written the potential risk as a function of the observed data, we are in need of one more assumption, known as **positivity**.[15] Positivity requires exposed and unexposed individuals within all confounding levels (Mortimer et al., 2005, Westreich and Cole (2010)). There are two kinds of positivity violations (non-positivity): structural (or deterministic) and stochastic[16] (or random).

[15] Also known as the experimental treatment assignment assumption.

[16] The word **stochastic** is derived from the greek word "to aim," as in "to aim for a target."

Structural non-positivity occurs when individuals with certain covariate values cannot be exposed. For example, in occupational epidemiology work-status (employed/unemployed in workplace under study) is a confounder, but individuals who leave the workplace can no longer be exposed to a work-based exposure. Alternatively, stochastic non-positivity arises when the sample size is not large enough to populate all confounder strata with observations.

Problems because of positivity arise for two reasons. The first is definitional. Consider the step in our equation above where we marginalize over C to equate the potential and observed outcomes. In the case where C is binary and we want to estimate the potential outcome if everyone were exposed to $X = 1$, this step could be re-written as:

$$E(Y^{x=1}) = E(Y \mid X = 1, C = 1)P(C = 1) + E(Y \mid X = 1, C = 0)P(C = 0)$$

Now imagine that for those with $C = 1$, it is either impossible to have $X = 1$ (structural nonpositivity) or we just don't have anyone in our sample with $X = 1$ (stochastic nonpositivity). Mathematically, it does not make sense to write $E(Y \mid X = 1, C = 1)$ because there are no individuals with $X = 1$ and $C = 1$. We thus cannot define this conditional average.

The second problem with positivity violations has to do with estimators. Consider, for example, a simple inverse probability weight that corresponds to the above scenario (i.e., if $C = 1$, there are no individuals with $X = 1$):

$$\frac{1}{P(X = 1 \mid C = 1)}$$

In this case, the probability in the denominator is zero. And because $1/0$ is undefined, we can't use IP-weighting to estimate the effect we're after with this estimator. The same type of problem arises even if there are only a very small number people in the sample with $X = 1$ if $C = 1$. In this latter case, imagine that the probability of being exposed is very small, say 0.0001. Then, the above weight would be equivalent to $1/0.0001 = 10,000$. The above weight means that one or more of these individuals will contribute $10,000$ observations to the weighted analysis (usually well more than the original sample). These types of problems result in instability of the estimator (because the results end up being heavily dependent on only a few individuals in the sample with large weights).

When faced with positivity violations, one should either re-define the estimand so that there is no positivity violation, choose an estimator that is less affected by positivity problems, or both (Petersen et al., 2012).[17] Alternative estimands include the effect of treatment on the treated or untreated, various types of stochastic effects [including incremental propensity score effects (Kennedy, 2019), which do not require that positivity hold (Naimi et al., 2021)], or "blip" effects that are encoded in structural nested models, and can be estimated with g estimation. One can also use collaborative targeted minimum loss-based estimation,[18] and the parametric g formula, which tend to be less sensitive to positivity violations (Cole et al., 2013; Porter et al., 2011; Ju et al.,

[17] Keep in mind: one cannot simply "avoid" positivity. In extreme setting, nonpositivity means that those who were unexposed in the sample are very unlikely to be exposed (and vice versa). In such a situation, it may not make sense to estimate the average treatment effect, because there is a subset of the population who may never realistically be exposed (or unexposed). In this case, g estimation, cTMLE, and the parametric g formula can actually estimate parameters that differ slightly or profoundly from the ATE.

[18] there is mounting evidence that standard (not collaborative) TMLE is very sensitive to positivity violations.

2017).

There are a number of different procedures one can use to evaluate whether positivity is a problem. Among these include propensity score overlap plots. Consider again our data from the last section. To get the propensity score for a binary exposure, we can fit a logistic model to the exposure data, conditional on confounders. Here, we use the Lalonde dataset, which is well known in econometric circles. This dataset was originally obtained from a study used to evaluate the effect of a training program (treat) on income:

```
library(MatchIt)
data("lalonde")

head(lalonde)
```

```
##      treat age educ   race married nodegree re74 re75       re78
## NSW1     1  37   11  black       1        1    0    0  9930.0460
## NSW2     1  22    9 hispan       0        1    0    0  3595.8940
## NSW3     1  30   12  black       0        0    0    0 24909.4500
## NSW4     1  27   11  black       0        1    0    0  7506.1460
## NSW5     1  33    8  black       0        1    0    0   289.7899
## NSW6     1  22    9  black       0        1    0    0  4056.4940
```

```
propensity_score <- glm(treat ~ age + educ +
    re74 + re75, data = lalonde, family = binomial(link = "logit"))$fitted.values

head(propensity_score)
```

```
##      NSW1      NSW2      NSW3      NSW4      NSW5      NSW6
## 0.3916080 0.3824366 0.4084567 0.3999990 0.3627466 0.3824366
```

```
## by appending a '$fitted.values' to
## the end of this glm function, we are
## keeping the predicted values from
## the model under the observed data
## settings.
```

We can now plot the density of this propensity score for each exposure group to see how they overlap:

```
set.seed(123)

exposure <- lalonde$treat

plot_data <- data.frame(propensity_score,
    Exposure = as.factor(lalonde$treat))

p1 <- ggplot(data = plot_data) + scale_y_continuous(expand = c(0,
    0)) + scale_x_continuous(expand = c(0,
    0)) + ylab("Density") + xlab("Propensity Score") +
    scale_color_manual(values = c("#000000",
        "#D55E00")) + geom_density(aes(x = propensity_score,
    group = Exposure, color = Exposure)) +
    geom_histogram(aes(y = ..density.., x = propensity_score,
        alpha = 0.25, group = Exposure, color = Exposure)) +
    xlim(0, 1)

ggsave(here("figures", "2022_01_10-ps_overlap.pdf"),
    plot = p1)
```

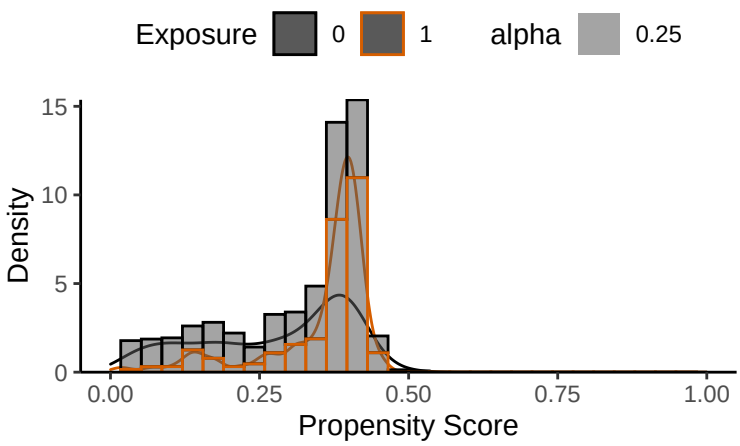

Figure 1: Propensity score overlap plot for the training intervention in 614 individuals in the Lalonde dataset.

Since the mass of the density for the exposed occurs in the same place as

the density mass for the unexposed, positivity does not seem to be much of an issue here. Another way to check positivity is to create stabilized inverse probability weights[19] and look at their descriptive statistics.

[19] We won't get too deep into the theory for / definition of weights here. But here is some code for creating stabilized weights and evaluating positivity.

```
sw <- (mean(exposure)/propensity_score) *
    exposure + ((1 - mean(exposure))/(1 -
    propensity_score)) * (1 - exposure)

summary(sw)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6548  0.7788  0.9340  1.0451  1.1245 27.6692
```

The mean of the stabilized weights is 1, and the max weight is not large at all, suggesting very well-behaved weights. Thus, in this particular case, we are not concerned with violations of the positivity assumption.

Technical Note:

In a large body of methods literature, particularly econometrics, you are likely to encounter these causal assumptions articulated in different ways. Most commonly, researchers will often invoke **ignorability** as a core assumption in causal inference. There are at least two versions of ignorability: strong and weak. **Strong ignorability** is defined as the combination of the conditional independence assumption, and the positivity assumption. Technically, strong ignorability holds if, for individual i with a binary exposure $X \in \{0, 1\}$:

$$(Y_i^{x=0}, Y_i^{x=1}) \perp\!\!\!\perp X_i \mid \mathbf{C}_i, \text{ and}$$
$$0 < P(X_i = 1 \mid \mathbf{C}_i) < 1,$$

where the $\perp\!\!\!\perp$ symbol denotes independence (in this case, conditional independence since we include $\mid \mathbf{C}_i$). In this case, the ignorability is "strong" because the independence is assumed to exist *jointly* between both potential outcomes (Y_i^{x=0}, Y_i^{x=1}) for individual i, and the exposure X_i, conditional on $\mathbf{C}_i$. Sometimes, a weaker version of the assumption is made:

$$(Y_i^{x}) \perp\!\!\!\perp X_i \mid \mathbf{C}_i, \text{ and}$$
$$0 < P(X_i = 1 \mid \mathbf{C}_i) < 1,$$

This version of the assumption is weaker in that we need not worry about whether the potential outcomes are jointly independent of the observed exposure. Rather, we only need each potential outcome (Y_i^x) to be independent of the observed exposure.

Even still, these assumptions are stronger than what we need to identify the causal risk difference, risk ratio, odds ratio, or other typical summary contrasts we often quantify in epidemiology. In the above proof, we demonstrated that identifiability is obtained under **mean exchangeability** $E(Y^x \mid X, \mathbf{C}) = E(Y^x \mid \mathbf{C})$. This assumption is even weaker than those articulated in the formalization of strong and weak ignorability.

The distinctions between strong ignorabilty, weak ignorability, and mean exchangeability are of little practical consequence. While it is important to recognize that ignorability typically consists of the combination of exchangeability and positivity. Additional details on and intuition behind the different versions of exchangeability can be found in Technical Point 2.1 of Hernán and Robins (2020).

6 What now? Choosing the Estimator

Consider the example above, where we had to adjust for C to equate the potential and observed outcomes. In our simple example, we only considered one confounding variable (sleep quantity), but in a typical observational study, we'd adjust for quite a few variables. Consider further that we'd typically employ a statistical regression model (e.g., linear, logistic, Poisson, Cox, or other) to actually implement our adjustment, which might look something like[20]:

[20] Such a model would be what we'd use in SAS, Stata, R, or any other software when we use the regression function and include only main effects terms in the model

$$E(Y \mid X, C_1, C_2, C_3, C_4) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4$$

The problem with using the above model is that it makes fairly strong assumptions about exactly *how* Y is related to X and the confounders. Specifically, this equation states (or assumes) that the conditional mean of Y is related to all the variables additively such that a single unit increase in each variable results in a linear and independent increase in the mean of Y.

However, consider that for five variables there can be a total of[21]

[21] This equation is referred to as the binomial coefficient.

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$$

two-way interactions that we could potentially add to the model. Additionally, we could include higher-order interactions, for example, a three-way interaction between X, C_1, and C_3. In fact, if we considered higher order interactions, for this simple model would could have up to:

$$2^5 - 5 - 1 = 26$$

k-way interactions (including 2, 3, 4, and 5 way). If we exclude any of the relevant interactions from among this set, our model would be misspecified. This misspecification could result in bias.

There are many other choices that can lead to problems with the estimator, including making linearity (or nonlinearity) assumptions, choosing the link functions in a generalized linear model (see, e.g., Weisberg and Welsh, 1994), or making the distributional assumption about the conditional mean of the outcome. It is for these reasons (among others) that machine learning methods are becoming so popular. Generally, machine learning methods do not tend

to rely as heavily on such (parametric) assumptions about how the data were generated. However, they do come with some important trade-offs that should be considered before use. We will consider these tradeoffs shortly.

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *J Am Stat Assoc*, 91(434): 444–455, 1996.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition, 2002.

Stephen R. Cole, David B. Richardson, Haitao Chu, and Ashley I. Naimi. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. *Am J Epidemiol*, 177(9):989–996, 2013.

Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, New Haven, CT, 2021.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Sander Greenland and James Robins. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov*, 6(1):4, 2009.

Sander Greenland and JM Robins. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*, 15(3):413–419, 1986.

Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Stat Sci*, 14(1):29–46, 1999.

M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: A recent methodological review. *Curr Epidemiol Rep*, 3(4):297–305, Dec 2016.

M. A. Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

M A Hernan and S L Taubman. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *Int J Obes*, 32(S3): S8–S14, 2008.

Miguel A. Hernán. Invited commentary: Hypothetical interventions to define causal effects—afterthought or prerequisite? *Am J Epidemiol*, 162(7):618–620, 2005.

Miguel A Hernán and Tyler J VanderWeele. Compound treatments and transportability of causal inference. *Epidemiol*, 22(3):368–377, May 2011. DOI: 10.1097/EDE.0b013e3182109296.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *J Am Stat Assoc*, 103(482):832–842, 2008.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, 2015.

Cheng Ju, Susan Gruber, Samuel D Lendle, Antoine Chambaz, Jessica M Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J van der Laan. Scalable collaborative targeted learning for high-dimensional data. *Statistical Methods in Medical Research*, 28(2):532–554, 2017.

Edward H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Kathleen M Mortimer, Romain Neugebauer, Mark van der Laan, and Ira B Tager. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol*, 162(4):382–388, Aug 2005. DOI: 10.1093/aje/kwi208.

Ivan Diaz Munoz and Mark van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

A I Naimi. The Counterfactual Implications of Fundamental Cause Theory. *Curr Epidemiol Reports*, In Press, 2016.

Ashley I. Naimi and Jay S. Kaufman. Counterfactual theory in social epidemiology: Reconciling analysis and action for the social determinants of health. *Curr Epidemiol Reports*, 2(1):52–60, 2015.

Ashley I. Naimi and Daniel J. Westreich. Big data: A revolution that will transform how we live, work, and think. *American Journal of Epidemiology*, 179(9):1143–1144, 2014.

Ashley I Naimi and Brian W Whitcomb. Estimating risk ratios and risk differences using regression. *American Journal of Epidemiology*, 189(6):508–510, 2020.

Ashley I. Naimi, E Rudolph, H Kennedy, A Cartus, SI Kirkpatrick, DM Haas, H Simhan, and LM Bodnar. Incremental propensity score effects for time-fixed exposures. *Epidemiology*, 32(2):202–208, 2021.

Judea Pearl, Madelyn R Glymour, and Nicholas Jewell. *Causal Inference in Statistics: A Primer*. Wiley, United Kingdom, 2016.

Karl Pearson. *The Grammar of Science*. London, J.M. Dent & sons ltd, 3rd edition, 1911.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Stat Methods in Med Res*, 21(1):31–54, 2012.

Kristin E Porter, Susan Gruber, Mark J van der Laan, and Jasjeet S Sekhon. The relative performance of targeted maximum likelihood estimators. *Int J Biostat*, 7(1), 2011.

James M. Robins and Sander Greenland. Adjusting for differential rates of prophylaxis therapy for pcp in high-versus low-dose azt treatment arms in an aids randomized trial. *J Am Stat Assoc*, 89(427):737–749, 1994.

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *J Am Stat Assoc*, 75(371):591–593, 1980.

Donald B Rubin. Causal inference using potential outcomes. *J Am Stat Assoc*, 100(469):322–331, 2005.

Ian Shrier, Russell J Steele, Evert Verhagen, Rob Herbert, Corinne A Riddell, and Jay S Kaufman. Beyond intention to treat: what is the right question? *Clin Trials*, 11(1):28–37, Feb 2014. DOI: 10.1177/1740774513504151.

Eric J. Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Stat Med*, 33(21):3601–3628, 2014.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Stat Methods in Med Res*, 21(1):55–75, 2012.

Tyler J VanderWeele and Miguel Ángel Hernán. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.

S. Weisberg and A. H. Welsh. Adapting for the missing link. *The Annals of Statistics*, 22(4):1674—1700, 1994.

Daniel Westreich and Stephen R. Cole. Invited commentary: Positivity in practice. *Am J Epidemiol*, 171(6):674–677, 2010.