Biostatistics Lecture Notes

Causal Inference

Ming-Chieh Shih

June 6, 2023

In statistical modelling, we are often told that "*Correlation does not imply causation*", and are given examples like the association between the crowing of roosters and sunrise, or the association between ice cream sales and rate of drowning. Some will stop here and just state that we have to treat associations "carefully", and some will come up with a set of common rules to detect non-causal associations (eg. look for temporality to avoid reverse causation, or look for common cause that leading to confounding). These rules are not wrong, but they do not tackle the *correlation does not imply causation* dilemma in a systematic way. That is, we need to know *what set of assumptions we need to make* before we can tie correlation between treatment A and outcome B to the actual causal effect of treatment A on B. Here, the treatment A we are referring to is not limit to medical treatments, and can be any characteristic or action we ("or God") can manipulate, e.g. promotion of health policies, alternating sales strategies, changing dieting habits, etc.

The aim of "*causal inference*" is to study the mathematical formulation, assumption requirement, and estimation strategies for causal effects. In this course, we will follow the content of the book *Causal Inference: What If* by Miguel Hernán and James Robins (available here) and go through some of the basic tools for modern causal inference.

1 Counterfactuals and Causal Effects

We start from the concept of counterfactuals. Suppose that like in the textbook, a patient i waiting for heart transplant received a new heart, and then died a year later. Suppose we denote the indicator random variable for "death within one year" as Y, the the indicator random variable for receiving a heart transplant as A. For now, we assume that the outcome is deterministic. Then for patient i, we have

$$Y_i = 1, \quad A_i = 1$$

However, from this information alone we do not know if and how the surgery affected patient i's survival. Whether the causal effect exists depends on which of the following is true:

a) If patient i would have died anyway, with or without the surgery, then the surgery did not affect their survival.

b) If patient i would have lived if they did not receive the surgery, then the surgery seems to have caused their death.

Therefore, before we could say any about the causal effect of the surgery on patient i's survival, we will need to know their survival *had they not received a heart transplant.*

To cater to this query, we define a vector of random variables $(Y^{a=0}, Y^{a=1})^\top$, or shorthanded as $(Y^0, Y^1)^\top$, that stands for the *counterfactual* outcomes had the patient received different treatment decisions: Y^0 is their survival had they *not* received the surgery, and Y^1 is their survival had they received the surgery. For patient i, we have already known that receiving the surgery results in death within one year, so we should have $Y_i^1 = 1$. This intuition is formalized as the **consistency assumption**:

Assumption 1 (Consistency) $Y = Y^A$.

In our query for the causal effect of the surgery on patient i, for scenario a) above, we have $Y_i^0 = 1$, so that $Y_i^0 = Y_i^1 = 1$ and there is no *individual causal effect* on patients i. For scenario b) above, we have $Y_i^0 = 0$, so that $Y_i^0 \neq Y_i^1$ and the surgery has *individual causal effect* on patient i. This leads to the formal definition

Definition 1 (Individual causal effect) *Treatment A has a causal effect on individual i if and only if $Y_i^1 \neq Y_i^0$.*

In reality, we cannot observe Y_i^0 because patient i actually received the surgery, rendering it *counterfactual* and thus the name for $(Y^0, Y^1)^\top$. Since we usually cannot identify the individual causal effect, we turn to the population causal effect, which compares the distribution of Y^1 and Y^0 within a certain population. In particular, we are often interested in whether the *mean* of Y^1 and Y^0 are different. The mean of Y^1 is the average outcome of the population of interest if everyone received the surgery, and the mean of Y^0 is the average outcome when no-one received the surgery. Therefore, when the two means are different, it implies that on average, the treatment has a causal effect on the outcome, vice versa. In other words, the *null hypothesis of no average causal effect* would be

$$\mathbb{E}[Y^1] = \mathbb{E}[Y^0] \tag{1}$$

where the expectation is taken over our population of interest. For example, suppose our population is finite with Y_i^0 and Y_i^1 as the following left panel:

ID	Y^0	Y^1
1	0	1
2	1	0
3	0	0
4	0	0
5	0	0
6	1	0
7	0	0
8	0	1
9	1	1
10	1	0
11	0	1
12	1	1
13	1	1
14	0	1
15	0	1
16	0	1
17	1	1
18	1	0
19	1	0
20	1	0

Then we have $\mathbb{E}[Y^0] = \mathbb{E}[Y^1] = 0.5$, so there is no average causal effect. However, notice that even if there is no average causal effect, the treatment still has causal effect on certain individuals, eg. patients with ID 1 and 2. No individual causal effect for all individuals implies no average causal effect, but the reverse does not hold.

The null hypothesis of average causal effect cares about the equivalence of $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$, so we can define the measure of causal effect by some functions of $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$ that takes on a specific value when $\mathbb{E}[Y^1] = \mathbb{E}[Y^0]$ and other values otherwise. The most commonly used measure of causal effect in causal inference is the mean difference:

$$\tau = \mathbb{E}[Y^1] - \mathbb{E}[Y^0] \tag{2}$$

When Y^1 and Y^0 are binary, we have

$$\tau_{\text{RD}} = \mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{P}[Y^1 = 1] - \mathbb{P}[Y^0 = 0] \tag{3}$$

which translates to the *risk difference* between the two treatment decisions. Here $\tau_{\text{RD}} = 0$ implies no average causal effect. Alternatively, we can define the following measures of causal effect for binary outcomes

$$\tau_{\text{RR}} = \frac{\mathbb{P}[Y^1 = 1]}{\mathbb{P}[Y^0 = 1]} \tag{4}$$

$$\tau_{\text{OR}} = \frac{\mathbb{P}[Y^1 = 1]/\mathbb{P}[Y^1 = 0]}{\mathbb{P}[Y^0 = 1]/\mathbb{P}[Y^0 = 0]} \tag{5}$$

Equation (4) defines the causal effect as the ratio of outcome risks between the two treatment decisions, and is termed as *risk ratio* or *relative risk* by epidemiologists. The value of risk ratio under the null hypothesis of no causal effect is 1. Equation (5) defines the causal effect as the ratio of outcome odds between the two treatment decisions, and is termed as *odds ratio* by epidemiologists. The value of odds ratio under the null hypothesis of no causal effect is also 1.

In the previous example, we assume that our population of interest is finite and only consists of 20 patients. Therefore, if we can gather their counterfactual outcomes, we can calculate the exact measures of causal effect such and risk difference or risk ratio. However, in reality, the population of interest (eg. all heart failure patients in Taiwan) is not observable, and we can only obtain a sample of the population. Therefore, we would have to estimate $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$ using the sample, and the most intuitive way is to use the *empirical mean* defined as

$$\mathbb{E}_n[Y^1] = \frac{1}{n}\sum_{i=1}^{n} Y_i^1 \tag{6}$$

$$\mathbb{E}_n[Y^0] = \frac{1}{n}\sum_{i=1}^{n} Y_i^0 \tag{7}$$

where n is the sample size. As long as the individuals are mutually independent and Y^1, Y^0 have finite variance, the empirical means are *consistent* estimators for $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$, so we can estimate the risk difference, risk ratio and odds ratio as:

$$\hat{\tau}_{\text{RD}} = \mathbb{P}_n[Y^1 = 1] - \mathbb{P}_n[Y^0 = 1] \tag{8}$$

$$\hat{\tau}_{\text{RR}} = \frac{\mathbb{P}_n[Y^1 = 1]}{\mathbb{P}_n[Y^0 = 1]} \tag{9}$$

$$\hat{\tau}_{\text{OR}} = \frac{\mathbb{P}_n[Y^1 = 1]/\mathbb{P}_n[Y^1 = 0]}{\mathbb{P}_n[Y^0 = 1]/\mathbb{P}_n[Y^0 = 0]} \tag{10}$$

where $\mathbb{P}_n[W = 1] := \mathbb{E}_n[W]$ and $\mathbb{P}_n[W = 0] := \mathbb{E}_n[1 - W]$ are the empirical proportions for $W = 1$ and $W = 0$ given that W is binary. These estimators would also be consistent for their corresponding effect measures τ_{RD}, τ_{RD} and τ_{RD}.

2 Causation, Association and Marginal Exchangeability

ID	A	Y	Y^0	Y^1
1	0	0	0	1
2	0	1	1	0
3	0	0	0	0
4	0	0	0	0
5	1	0	0	0
6	1	0	1	0
7	1	0	0	0
8	1	1	0	1
9	0	1	1	1
10	0	1	1	0
11	0	0	0	1
12	1	1	1	1
13	1	1	1	1
14	1	1	0	1
15	1	1	0	1
16	1	1	0	1
17	1	1	1	1
18	1	0	1	0
19	1	0	1	0
20	1	0	1	0

Previously we showed how to use the empirical mean of counterfactual outcomes to estimate the average causal effect under selected effect measures. However, in reality we only observe one of Y^0 and Y^1 for each individual, as shown in the left panel. When subject i did not receive the surgery, we observe $Y_i = Y_i^0$ from the consistency assumption, and Y_i^1 is unknown, or in a sense missing (shown in gray). When subject i did receive the surgery, we observe $Y_i = Y_i^1$, and Y_i^0 is missing. In this case, the best we can do to estimate $\mathbb{E}[Y^1]$ is to compute the empirical mean of observable Y^1, which is the Y for those $A = 1$. The same is done for $\mathbb{E}[Y^0]$. Therefore, we are using the following estimators for $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$:

$$\hat{\mathbb{E}}[Y^1] = \mathbb{E}_n[Y|A = 1] \tag{11}$$

$$\hat{\mathbb{E}}[Y^0] = \mathbb{E}_n[Y|A = 0] \tag{12}$$

Then under binary Y, the risk difference, risk ratio and odds ratio estimators would be

$$\hat{\tau}_{\text{RD}} = \mathbb{P}_n[Y = 1|A = 1] - \mathbb{P}_n[Y = 1|A = 0] \tag{13}$$

$$\hat{\tau}_{\text{RR}} = \frac{\mathbb{P}_n[Y = 1|A = 1]}{\mathbb{P}_n[Y = 1|A = 0]} \tag{14}$$

$$\hat{\tau}_{\text{OR}} = \frac{\mathbb{P}_n[Y = 1|A = 1]/\mathbb{P}_n[Y = 0|A = 1]}{\mathbb{P}_n[Y = 1|A = 0]/\mathbb{P}_n[Y = 0|A = 0]} \tag{15}$$

To derive the necessary assumptions for these estimator to be consistent, let us ignore the counterfactuals for a bit and look at the causal effect estimator for risk difference in Equation (13). Since $\mathbb{P}_n$ is an empirical mean, $\hat{\tau}_{\text{RD}}$ is consistent for the association measure

$$\theta_{\text{RD}} = \mathbb{P}[Y = 1|A = 1] - \mathbb{P}[Y = 1|A = 0] \tag{16}$$

We say that θ is an association measure because it measures the strength of association between Y and A. In addition, when $\theta_{\text{RD}} = 0$, Y and A are independent, or $Y \perp\!\!\!\perp A$. To see this, we have, with binary Y and A,

$$Y \perp\!\!\!\perp A \tag{17}$$

$$\Leftrightarrow \quad \mathbb{P}[Y = y|A = a] = \mathbb{P}[Y = y] \qquad \forall y, a \in \{0, 1\} \tag{18}$$

$$\Leftrightarrow \quad \mathbb{P}[Y = y|A = 1] = \mathbb{P}[Y = y|A = 0] \qquad \forall y \in \{0, 1\} \tag{19}$$

$$\Leftrightarrow \quad \mathbb{P}[Y = 1|A = 1] = \mathbb{P}[Y = 1|A = 0] \tag{20}$$

$$\Leftrightarrow \quad \mathbb{P}[Y = 1|A = 1] - \mathbb{P}[Y = 1|A = 0] = 0 \tag{21}$$

Therefore, $\hat{\tau}_{\text{RD}}$ is estimating the association between A and Y, represented by $\mathbb{P}[Y = 1|A = 1] - \mathbb{P}[Y = 1|A = 0]$. However, what we actually want is to estimate the average causal effect of A on Y, i.e. $\tau_{\text{RD}} = \mathbb{P}[Y^1 = 1] - \mathbb{P}[Y^0 = 1]$. A sufficient condition that makes these two quantities equivalent is the *exchangeablity* assumption:

Assumption 2 (Exchangeablity) $Y^a \perp\!\!\!\perp A, \forall a \in \{0, 1\}$.

To see this, under the consistency and exchangeablity assumptions, for all $a \in \{0, 1\}$:

$$\begin{aligned} \mathbb{P}[Y = 1|A = a] &= \mathbb{P}[Y^a = 1|A = a] && \text{(Consistency)} && (22) \\ &= \mathbb{P}[Y^a = 1] && \text{(Exchangeablity)} && (23) \end{aligned}$$

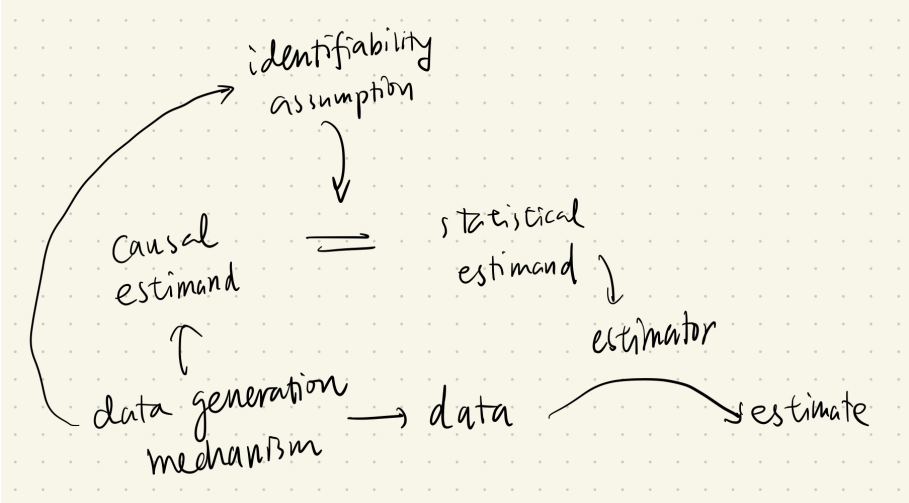
Therefore we have,

$$\tau_{\text{RD}} = \mathbb{P}[Y^1 = 1] - \mathbb{P}[Y^0 = 1] = \mathbb{P}[Y = 1|A = 1] - \mathbb{P}[Y = 1|A = 0] \tag{24}$$

Note the in Equation (24), the left hand side is a counterfactual quantity that represents the causal effect of A, which we term as the *causal estimand*. The causal estimand is not readily estimable since Y^a is partially missing. The right hand side is a quantity estimable by data that represents the association between A and Y, which we term as the *statistical estimand*. Assumptions 1 and 2 are the assumptions that ties these two estimands together so that the causal effect is estimable from data, i.e. *identifiable*. Therefore, Assumptions 1 and 2 are often called the *identifiability assumptions*. The recipe for estimating the statistical estimand, $\mathbb{P}_n[Y = 1|A = 1] - \mathbb{P}_n[Y = 1|A = 0]$ is called the *estimator*, and the number produced by the estimator using data is called the *estimate* of causal effect (under the selected measure of causal effect). In the data above, the causal effect estimate using risk difference as effect measure would be

$$\hat{\tau}_{\text{RD}} = \mathbb{P}_n[Y = 1|A = 1] - \mathbb{P}_n[Y = 1|A = 0] = \frac{7}{13} - \frac{3}{7} = \frac{10}{91} \tag{25}$$

which should be tested against 0 taking the variation of the estimator into account. In summary, the roadmap of the causal inference framework is as follows:



Now let us dig deeper into Assumption 2, the exchangeability assumption:

$$Y^a \perp\!\!\!\perp A, \quad \forall a \in \{0, 1\} \tag{26}$$

which states that the assignment of the treatment A does not depend on the counterfactual outcomes of the patients. A scenario where this assumption holds is simple randomization in randomized experiments, where each patient is randomized to heart transplant with probability p and no heart transplant with probability $1 - p$, irrelevant of their baseline characteristics. In this case, we have for any $a \in \{0, 1\}$

$$\mathbb{P}[A = 1|Y^a] = \mathbb{P}[A = 1] = p \tag{27}$$

since the treatment assignment is by design simply flipping a (biased) coin, and should be indepedent of Y^a. Given that A is binary, this condition is equivalent to $Y^a \perp\!\!\!\perp A$, and thus the exchangeability assumption is satisfied. Since in this exchangeability assumption, A and Y^a is

independent over the whole population, we also call it the *marginal exchangeability assumption*. Later we will see that we can relax this assumption into *conditional* exchangeability, at the cost of slightly more complicated statistical estimand.

Violation of the marginal exchangeability assumption is common in biomedical or epidemiological studies, since the treatment is often non-randomized. In the case of heart transplants, physicians and policy makers will tend to assign donated hearts to those who are expected to have better prognosis and less complication after receiving the heart transplant, eg. younger patients and patients without too much underlying diseases. In this case, $\mathbb{P}[Y^1 = 1|A = 1]$ is expected to be smaller than $\mathbb{P}[Y^1 = 1|A = 0]$, which implies that Y^1 is not independent to A, thus violating the marginal exchangeability assumption. From an estimand construction standpoint, originally we would want to replace the causal estimand $\mathbb{E}[Y^1]$ by $\mathbb{E}[Y|A = 1]$. However, under this treatment assignment regime, the actual average death rate when every patient is sent for surgery ($\mathbb{E}[Y^1]$) would be higher than that observed in the data ($\mathbb{E}[Y|A = 1]$), therefore the replacement is not valid. In later sections, we will see how we can adjust for this violation of marginal exchangeablity by assuming the weaker conditional exchangeability.

3 ATE under Conditional Exchangeability

In the previous section, we assumed that the treatments were randomized between the two treatments of our interest, so that we have marginal exchangeability of counterfactual outcomes between the two treatment groups. However, often times, especially in observational studies, the assignment of treatments are determined by decision makers, eg. physicians and policy enforcers, and their choice among the treatments are usually guided by some other covariates. For example, for physicians, the frailty and prognosis of the patients may play an important role in the treatment they are going to receive. Therefore, for most types of data, we need a way to relax the requirement of marginal exchangeability to be still able to estimate the average treatment effects (ATEs).

ID	L	A	Y
1	0	0	0
2	0	0	1
3	0	0	0
4	0	0	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	1
9	1	0	1
10	1	0	1
11	1	0	0
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	0
19	1	1	0
20	1	1	0

The idea for relaxing marginal exchangeability comes from the analysis of *conditionally randomized experiments*. In conditionally randomized experiments, the participants are first grouped by their characteristics, eg. sex as male or female. Then, for each group of patients, we randomize them to either treatment with predefined probabilities, but the these probabilities can be different between groups. For example, in the table on the left, suppose L represents the priority on the list for transplantation, taking value 1 for high-priority and 0 for low-priority. Then the probability for low-priorities receiving transplant (i.e. treatment 1) is $\frac{4}{8} = 0.5$, while the probability for high-priorities receiving transplant is $\frac{9}{12} = 0.75$. However, within the high- and low-priority groups, the treatments are still randomized, i.e. their treatments decided by a flip of biased coin with probability of being 1 as 0.5 for low-priorities and 0.75 for high-priorities. In this case, we say that **conditional exchangeability** holds, which is defined as follows.

Assumption 3 (Conditional exchangeability) $Y^a \perp\!\!\!\perp A|L, \forall a \in \{0, 1\}$

Note that under conditional exchangeability (and consistency), we still cannot estimate the ATE using mean difference between the two treatments, since the average counterfactual outcome

for treatment a can be written as

$$\mathbb{E}[Y^a] = \sum_l \mathbb{E}[Y^a|L=l]\mathbb{P}[L=l] \tag{28}$$

$$= \sum_l \mathbb{E}[Y^a|A=a, L=l]\mathbb{P}[L=l] \qquad \text{(Conditional exchangeability)} \tag{29}$$

$$= \sum_l \mathbb{E}[Y|A=a, L=l]\mathbb{P}[L=l] \qquad \text{(Consistency)} \tag{30}$$

where in the summation l goes though all possible values for L. However, the mean outcome for participants receiving treatment a is

$$\mathbb{E}[Y|A=a] = \sum_l \mathbb{E}[Y|A=a, L=l]\mathbb{P}[L=l|A=a] \tag{31}$$

Therefore, when $\mathbb{P}[L=l] \neq \mathbb{P}[L=l|A=a]$, which is the case when the randomization probability is different across groups, Equation (30) and Equation (31) are usually not equivalent (An exception is when $\mathbb{E}[Y^a] \perp\!\!\!\perp L$, i.e. L does not have any effect on the prognosis). This is reasonable intuitively, since with conditional randomization with respect to priorities, the two treatment groups will then have different high-low priority proportions, and their mean outcome would thus be non-comparable.

Although we cannot estimate the ATE directly under only conditional exchangeability, we *can* estimate the **conditional average treatment effect** (CATE) with simple averages. The CATE (using mean outcome difference / risk difference as measure) is defined as $\tau(L) = \mathbb{E}[Y^1 - Y^0|L]$, i.e. the ATE *within* all patients with characteristic L. Under conditional exchangeability and consistency, from Equations (28) to (30), we have

$$\mathbb{E}[Y^a|L] = \mathbb{E}[Y|A=a, L] \tag{32}$$

Therefore, we can consistently estimate $\tau(L)$ with the mean outcome difference between the two treatment groups within the subpopulation with covariate value L, i.e. $\mathbb{E}[Y|A=1, L=0] - \mathbb{E}[Y|A=0, L=0]$. This is also intuitively correct, since within each level (or stratum) of L, we are still essentially doing a marginally randomized experiment, so calculating the mean difference of the two treatment groups would reflect the CATE within that stratum of L. This action of computing the stratum-specific treatment effects is terms as **stratification**. If we find that the CATE varies among strata, we say that the effect of the treatment is *modified* by L, or there is **effect modification** by L. For example, in the data above, we can estimate the CATE among low-priorities by

$$\mathbb{E}[Y|A=1, L=0] - \mathbb{E}[Y|A=0, L=0] = 1/4 - 1/4 = 0 \tag{33}$$

and the CATE among high-priorities can be estimated by

$$\mathbb{E}[Y|A=1, L=1] - \mathbb{E}[Y|A=0, L=1] = 6/9 - 2/3 = 0 \tag{34}$$

Therefore, the estimated CATE is identical between the two priority groups, and there is no evidence of effect modification by priority score.

We are now able to estimate the CATE for every stratum, but usually we are more interested in the ATE, since (1) in future application scenarios, we may not have access to the strata information for each patient (2) when the number of strata is large, which is especially true when there are large amounts of covariates, each individual CATE will be based on very small amount of data and thus very imprecise and impossible to do meaningful inference. Therefore, we need to find a way to leverage on the consistent CATE estimates to reconstruct an estimate for the ATE. There are two main approaches to achieve this: standardization and inverse probability weighting.

Standardization

In Equation (30), we showed that under conditional exchangeability with respect to L (and consistency), the mean counteractual outcome under treatment a can be written as a weighted average of estimated conditional outcomes

$$\mathbb{E}[Y^a] = \sum_l \mathbb{E}[Y|A=a, L=l]\mathbb{P}[L=l] \tag{35}$$

so the average treatment effect can also be written as a weighted average of CATEs

$$\mathbb{E}[Y^1 - Y^0] = \sum_l (\mathbb{E}[Y|A=1, L=l] - \mathbb{E}[Y|A=0, L=l])\mathbb{P}[L=l] \tag{36}$$

This method is called *standardization* in epidemiology. The reason standardization is important in epidemiology is that, say we want to compare the death rates of two different countries to reflect their medical standards. However, country A's citizens are primarily 20-50 years old, while country B's citizens are primarily 60-80 years old. Then, even if the two countries have the same medical standards, the *crude* death rate of country B would be higher than country A since it has more elderlies. A way to fairly compare these two countries is to first calculate the age-specific death rates of country A, say, θ_i for citizens of age i. Then, we denote the proportion of citizens that is of age i in country B as p_i. Then, the crude death rate for country A when its population structure is the same as country B is $\sum_i \theta_i p_i$, and this number can be compared with the crude death rate of country B.

In Equation (36), we are actually doing standardization of ATE with respect to the distribution of L, only that in this formula we are just standardizing into the population structure we observe, hence the weights $\mathbb{P}[L=l]$. In fact, if we would like to know the ATE for another population with different distribution of L, we can denote the probability of $L=l$ in that population as p_l and estimate the ATE with

$$\sum_l (\mathbb{E}[Y|A=1, L=l] - \mathbb{E}[Y|A=0, L=l])p_l \tag{37}$$

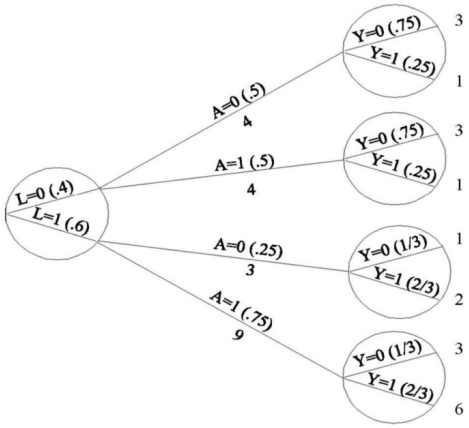
In our example data, since $P[L=0] = 12/20 = 0.6$ and $P[L=1] = 8/20 = 0.4$, we can estimate the ATE (using risk difference as measure) for the whole population by

$$\hat{\tau}_{\text{RD}} = \sum_{l=0}^{1} (\hat{\mathbb{E}}[Y|A=1, L=l] - \hat{\mathbb{E}}[Y|A=0, L=l])\hat{\mathbb{P}}[L=l] = 0 \cdot 0.6 + 0 \cdot 0.4 = 0 \tag{38}$$

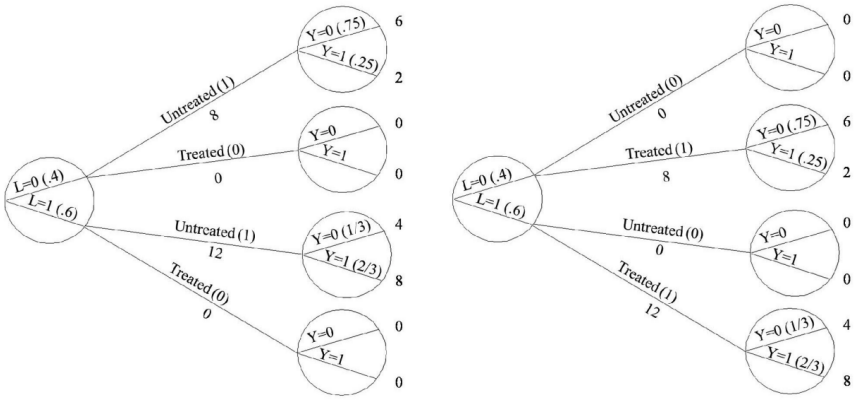
which is a little bit anticlimactic. However, in the case where there is effect modification, then the standardization would produce different results with different distribution of L.

Inverse probability weighting

Apart from standardization, we have another approach that can help us estimate the average treatment effect. First we see that our data can be visualized by the graph below: there are 8 patients of lower priority (with probability 0.4) and 12 patients of higher priority (with probability 0.6). Among the patients of lower priority, there is a probability of 0.5 receiving the transplant, while among the patients of higher priority, there is a probability of 0.75 receiving the transplant. The outcome is then drawn in the circles on the right hand side. In standardization, we first estimated the CATE for $L=0$ and $L=1$ by calculating the risk difference between the upper two circles and lower two circles (which are both 0), and then weighted these CATEs with the probability of $L=0$ (0.4) and $L=1$ (0.6). Notice that in this graph, if we calculated the crude risk difference by combining the second and fourth circle and contrasted it with the first and third circles, we would get a biased ATE estimate of $\frac{7}{13} - \frac{3}{7} \neq 0$.

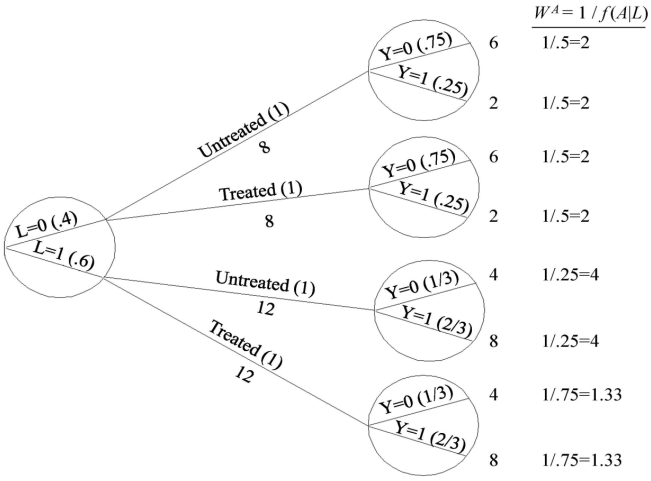


In inverse probability weighting, we are trying make comparing the crude risk difference work, but not without a little tweak to the data. First, let us imagine a counterfactual world where all patients are not treated. Since we have shown that $\mathbb{E}[Y|A=0,L]$ is equal to $\mathbb{E}[Y^0|L]$ under the conditional exchangeability assumption, we can use $\mathbb{E}[Y|A=0,L=0]=0.25$ and $\mathbb{E}[Y|A=0,L=1]=2/3$ to reconstruct the expected outcome, which is shown in the left panel in the figure below. Then, we can imagine another counterfactual world where all patients are treated. We can then use $\mathbb{E}[Y|A=1,L=0]=0.25$ and $\mathbb{E}[Y|A=1,L=1]=2/3$ to recreate the expected outcome under this world.



Now, we can combine the two panels above into one graph, shown as below. First notice that in this new "inflated" data, the treatment assignment is no longer related to L, since within each stratum of L, both treatments would have the same amount of patients as what that stratum originally had. In other words, the patients in this "inflated" data can actually be seen as randomly assigned to each treatment with probability 0.5. Another point to notice is that in this inflated data, the distribution of L is still $\mathbb{P}[L=1]=0.6$, which is the same as the original data. Therefore, we can treat this set of new data as randomized experiment and just compute the difference in mean outcome between the two treatment groups. In this case, it would be

$$\mathbb{E}_{\text{IPW}}[Y|A=1]-\mathbb{E}_{\text{IPW}}[Y|A=0]=\frac{10}{20}-\frac{10}{20}=0 \tag{39}$$



If we look back into how we derived the data in the counterfactual word, we can verify that a patient with treatment $A=a$ and covariate $L=l$ would be inflated by a factor of $\frac{1}{\mathbb{P}[A=a|L=l]}$, which is the inverse of the probability for that patient receiving his treatment. Therefore, this approach is called the **inverse probability weighting** approach. It can be proven that under non-parametric estimation of CATE and treatment probability, the standardization approach and the inverse probability weighting approach is actually equivalent.

4 Identifiability of ATE in Observational Studies

In the previous section, we showed that even if marginal exchangeability does not hold, as long as conditional exchangeability holds, along with the consistency assumption, we can identify the CATE using the data. In turn, we can use standardization or inverse probability weighting to identify the ATE. In the case where the variable L that is conditioned on is discrete, the statistical estimand that identifies the ATE (defined as mean difference / risk difference) via standardization, shown in Equation (36), is,

$$\tilde{\tau}^{(std)} = \sum_l (\mathbb{E}[Y|A=1, L=l] - \mathbb{E}[Y|A=0, L=l])\mathbb{P}[L=l] \tag{40}$$

which can be alternatively written using double expectations

$$\tilde{\tau}^{(std)} = \mathbb{E}_{L\sim F_L}\big[\mathbb{E}_Y[Y|A=1, L=l] - \mathbb{E}_Y[Y|A=0, L=l]\big] \tag{41}$$

Notice that the outer expectation is taken with respect to L, with L following the actual distribution of L for the population of interest, with cumulative density function denoted as F_L. This expression implies that even if L is continuous so that estimating $\mathbb{P}[L=l]$ in Equation (40) does not make sense, we can still make use of Equation (41) by the following procedure:

1. Obtain function estimates $\hat{\mu}_1(l)$ and $\hat{\mu}_0(l)$ for $\mathbb{E}[Y|A=1, L=l]$ and $\mathbb{E}[Y|A=0, L=l]$.
2. For *every* observation i within the population, calculate $\hat{\tau}_i^{(std)} = \hat{\mu}_1(L_i) - \hat{\mu}_0(L_i)$ and obtain the empirical average of $\hat{\tau}_i^{(std)}$ as the estimate for ATE:

$$\hat{\tau}^{(std)} = \frac{1}{N}\sum_i \hat{\tau}_i^{(std)} = \frac{1}{N}\sum_i [\hat{\mu}_1(L_i) - \hat{\mu}_0(L_i)] \tag{42}$$

For inverse probability weighting approach, the way we identified the ATE is to weight each individual with the reciprocal of the probability of them receiving their treatments base on their covariate values. Then, the mean counterfactual outcome $\mathbb{E}[Y^a]$ is estimated by averaging the outcomes of the "inflated" individual pool with $A=a$. However, it is not obvious how to average a pool of inflated individuals when the number of individuals is not integer after the weighting. Therefore, we now set out to formalize this weighting-averaging process with mathematical expressions. First notice that within the subpopulation with $A=a$, the standardized weight, i.e. the weight divided by the average weight within the subpopulation, is

$$w(L) = \frac{\frac{1}{\mathbb{P}[A=a|L]}}{\mathbb{E}_L\big[\frac{1}{\mathbb{P}[A=a|L]}\big|A=a\big]} \tag{43}$$

$$= \frac{\frac{1}{\mathbb{P}[A=a|L]}}{\mathbb{E}_L\big[\frac{f_L(L)}{f_{L|A}(L|A=a)\mathbb{P}[A=a]}\big|A=a\big]} \tag{44}$$

$$= \frac{\frac{1}{\mathbb{P}[A=a|L]}}{\frac{1}{\mathbb{P}[A=a]}\mathbb{E}_L\big[\frac{f_L(L)}{f_{L|A}(L|A=a)}\big|A=a\big]} \tag{45}$$

$$= \frac{\frac{1}{\mathbb{P}[A=a|L]}}{\frac{1}{\mathbb{P}[A=a]}\int \frac{f_L(l)}{f_{L|A}(l|a)} f_{L|A}(l|a)dl} \tag{46}$$

$$= \frac{\frac{1}{\mathbb{P}[A=a|L]}}{\frac{1}{\mathbb{P}[A=a]}\int f_L(l)dl} = \frac{\mathbb{P}[A=a]}{\mathbb{P}[A=a|L]} \tag{47}$$

Therefore, the statistical estimand for the ATE is

$$\tilde{\tau}^{(ipw)} = \mathbb{E}[w(L)Y|A=1] - \mathbb{E}[w(L)Y|A=0] \tag{48}$$

$$= \mathbb{E}\Big[\frac{Y\mathbb{P}[A=1]}{\mathbb{P}[A=1|L]}\Big|A=1\Big] - \mathbb{E}\Big[\frac{Y\mathbb{P}[A=0]}{\mathbb{P}[A=0|L]}\Big|A=0\Big] \tag{49}$$

$$= \mathbb{E}\Big[\frac{Y\mathbf{1}(A=1)}{\mathbb{P}[A=1|L]}\Big] - \mathbb{E}\Big[\frac{Y\mathbf{1}(A=0)}{\mathbb{P}[A=0|L]}\Big] \tag{50}$$

Therefore, based on Equation (50), we can obtain the inverse probability weighting estimator for ATE by the following procedure:

1. Obtain function estimates $\hat{e}_1(l)$ and $\hat{e}_0(l)$ for $\mathbb{P}[A=1|L=l]$ and $\mathbb{P}[A=0|L=l]$, which in the context of conditional randomized experiment are known and do not need to be estimated. $\hat{e}_1(l)$ is often termed as the **propensity score** among epidemiologists.

2. For *every* observation i within the population, calculate $\hat{\tau}_i^{(ipw)} = \big(\frac{\mathbf{1}(A_i=1)}{\hat{e}_1(L_i)} - \frac{\mathbf{1}(A_i=0)}{\hat{e}_0(L_i)}\big)Y$ and obtain the empirical average of $\hat{\tau}_i^{(ipw)}$ as the estimate for ATE:

$$\hat{\tau}^{(ipw)} = \frac{1}{N}\sum_i \hat{\tau}_i^{(ipw)} = \frac{1}{N}\sum_i \Big(\frac{\mathbf{1}(A_i=1)}{\hat{e}_1(L_i)} - \frac{\mathbf{1}(A_i=0)}{\hat{e}_0(L_i)}\Big)Y \tag{51}$$

In the previous section, we used the example of conditional randomization to illustrate a scenario where conditional exchangeability would hold. In the case of causal inference in observational data, the same set of assumptions and procedures can be carried over, with some caveats and amendments:

a) The conditional exchangeability assumption will be harder to justify: we are roughly assuming that all prognosis-related information that determines the treatment assignment is observed as covariates. In other words, under the medical setting, we are assuming that among the parameters the healthcare practitioner considered when making treatment decisions, we have measure all parameters that is related to the patient's future well-being. This is a very strong assumption since even if there is only one single unmeasured parameter, we are under serious risk of getting a biased ATE estimate.

b) For conditionally randomized experiments, the propensity scores are known by design. Yet for observational studies, the propensity scores have to be estimated.

c) In conditonally randomized experiments, the treatments are often given by experimenters, so the counterfactual outcomes Y^a defined in the consistency assumption is well-defined. However, for observational studies, the same treatment may have different versions among different subjects. For example, if the treatment is simply stated as "drug A", then different physicians may prescribe different doses, different intake intervals, and provide different levels of health education along with the prescription. In this case, Y^a is no longer well-defined and the ATE we are estimating may be fuzzy. What is worse is that, it may be the case that different versions of the treatments are not randomly distributed among subjects of different prognoses, which can introduce another layer of bias to our ATE estimate.

d) Notice that for the statistical estimand in standardization, we need the estimates of $\mathbb{E}[Y|A=1,L]$ and $\mathbb{E}[Y|A=0,L]$, which can be estimated non-parameterically only if we can observe individuals with $A=1$ and $A=0$ for any possible values of L. For example, if for any subpopulation with $L=l$, all individuals will be assigned to, say, $A=1$. Then, there is no way that we can find $\mathbb{E}[Y|A=0,L=l]$ unless we do some extrapolation. Similarly, for the inverse probability weighting estimator, we have $\mathbb{P}[A=1|L]$ and $\mathbb{P}[A=0|L]$ in the denominator, so they cannot be zero. In conditionally randomized experiments, the propensity score $\mathbb{P}[A=1|L]$ will always be designed to be between 0 and 1, i.e. participants within each strata of L will be randomized to either treatment, so the scenario above that breaks down causal inference would not occur. However, in observational studies, there is possibility that due to guidelines or clinical common sense, some types of patients will always receive one of the treatments. Therefore, we have to be careful that for these patients, we do not have the ability to infer their treatment effects, so we need to restrict our target population of inference in advance and remove those out of our scope. The assumption that ensures that the scenario above does not happen is the **Positivity** assumption, which is stated below

Assumption 4 (Positivity) $\mathbb{P}[A=a|L=l]>0,\ \forall a\in\{0,1\}, l\in\mathcal{X}_L$ *, where* $\mathcal{X}_L$ *is the support of* L *within the whole population.*

All in all, for observational studies, there are three basic assumptions we should make: *conditional exchangeability*, *consistency* and *positivity*, all with their own pitfalls. Under these assumptions, we can obtain (asymptotically) unbiased estimators for the ATE using the standardization approach or inverse probability weighting approach. The standardization approach focuses on estimating the relation between Y and A, L. The inverse probability weighting approach instead focuses on estimating the relation between A and L. Since these two estimators both provide (asymptotically) unbiased estimators while using different parts of the information from the data, it is possible to combine them to obtain a "merged" estimator that has smaller variance. We will look at this type of merged estimators in later sections.

5 Representation of Causal Structures using Causal Diagrams

In the previous sections, we talked about how conditional exchangeability could aid us in finding the ATE. In the easy case where there is only one covariate to be considered, it may be possible to argue conditional exchangeability from its definition. However, in the case where there are multiple variables with complex causal structures, it would then not be evident which identification strategy would be valid (eg. which variables should we put into the estimation of $\hat{\mu}_a$ during standardization). Fortunately, we can rely on a graphical representation of the causal structure, which we term as the **causal diagram**, to help us decide which variables to control for and which variables to avoid controlling for.

The simplest and most commonly used causal diagram is the causal directed acyclic graph (DAG), which connects variables with single arrows ("directed") that represents the causal relationship between variables. To prevent confusion of causal relationships, the directed graph should also be acyclic, which implies that any casual chains cannot form a loop, i.e. you can not have a chain with $A \Rightarrow B \Rightarrow C \Rightarrow A$. In the case where the variables have longitudinal measurements, the variables have to be splitted into different versions representing their values at different time points to account for temporality-hinted causal directions. Within this graph, we say A is a *direct cause* or *parent* of B if A has an arrow pointed to B; and we say that B is a *descendant* of A if there exists a path formed by forward arrows that goes from A to B.

Aside form being directed and acyclic, in order for the causal DAG to reflect the full information between the variables, there are two additional requirements for a causal DAG to be valid:

1. The lack of an arrow from one variable A to another variable B implies that there is no direct causal effect of A on B.

2. All common causes between any pair of variables should be included (i.e. absence of such variables imply no common causes between the two variables), even if they are not measured.

If we have a valid causal graph, then we say that A has a causal effect on B if there is at least one path from A to B that is formed by forward arrows.

The directed acyclic graph depicts the causal structure of the data. Yet to link the causal structure to the actual data distribution, we will often need the **causal Markov assumption**, which states that *any variable V, when conditioned on its direct causes, is independent to its non-descendents*. Under this assumption, suppose $V = (V_1, V_2, ..., V_M)^\top$ is the vector of variables on the graph with values $v = (v_1, v_2, ..., v_M)^\top$, then the probability density of v can be factored as

$$f_V(v) = \prod_{j=1}^{M} f_{V_j|PA_j}(v_j|pa_j) \tag{52}$$

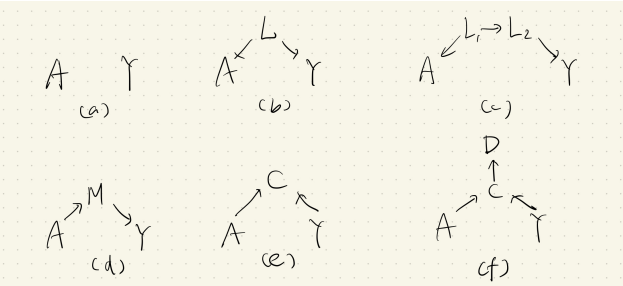
where pa_j is realized value for the vector of variables PA_j that are the direct causes of V_j. This is called the *Markov factorization* for the joint distribution of variables. For brevity, later we will omit the subscript for f and tell them part from their arguments.

The important value of causal DAGs and their implied Markov factorizations is that we can infer from the graph that if it implies there would be independence or conditional independence between variables. If a graph depicting absence of causal effect from A to Y implies (conditional) independence between them, but in data we detected a significant (conditional) association between A and Y, then we reject the DAG and conclude that there *is* causal effect, under the premise that the only alternative form of the graph is one with arrow from A to Y (i.e. the other part of the graph is assumed to be correct).

For example, the DAG in panel (a) above represents the causal structure of a marginally randomized experiment with absence of treatment effect. There are no common causes between the treatment A and outcome Y since the treatment is randomized. It is immediately clear that from the graph, A and Y should be marginally independent since from the Markov factorization:

$$f(A, Y) = f(A)f(Y) \tag{53}$$

where in the right hand side A and Y are not conditioned on any variable since they do not have direct causes. Therefore, in the case of marginally randomized experiments, suppose we see a



significant association between the treatment and the outcome, we can confidently say that this association is introduced by the causal (treatment) effect of the treatment on the outcome.

Panel (b) represents the causal structure of a conditional randomized experiment with L as the conditioned (vector) variable with no treatment effect for A on Y. This is also the DAG for observational studies where we have measured all common causes between A and Y that leads to conditional exchangeability. It can be shown that under this graph, A and Y are not marginally independent, but they are *conditionally* independent when conditioned on L. Therefore, when we are looking into every strata of L and realize that the treatment is actually associated with the outcome, this association cannot come from other places aside from the causal effect from A to Y that we originally thought to be absent.

For panel (c), L_2 is the only direct cause of outcome Y, while L_1 is the only direct cause for treatment A. In addition, L_1 is a direct cause of L_2. In this case, we can still show that A and Y are not marginally independent, but as long as you condition on either L_1 or L_2 (or event both!), A and Y will be conditionally independent, so detection of their conditional association would imply a casual effect from A to Y.

In the case of panel (d), there is a path from A to M to Y that is formed by forward arrows. Therefore, A should have a causal effect on Y. However, if we only looked at the conditional association between A and Y conditional on M, they would be conditionally independent. Notice that the absence of conditional association implied no causal effect of A on Y in panel (b), but the absence of conditional association here in panel (d) does not mean that A has no casual effect. Therefore, the way we interpret the conditional associations depends largely on the underlying causal DAG, and we should not willy-nilly condition on everything we've got, or we may risk missing a true causal effect.

For panels (e) and (f), A does not have a causal effect on Y (since the path from A to Y contains an arrow that is of the wrong direction). In this case, we can show that A and Y would be marginally independent, but conditional on C or D, A and Y be conditionally dependent. Therefore, here we have another case where conditioning on more variables does not do us good: we get the correct conclusion when we only look at the marginal association between A and Y, but arrive at the wrong conclusion when we conditional on either C or D.

The deduction of the previous graphs is intuitive and mathematically simple, but under a large causal DAG, it can be cumbersome to start from the Markov factorization every time. Fortunately, previous literature investigating the mathematical theory of the link between causal DAGs and implied (conditional) independence, which is called **d-separation**, has granted us a simple set of rules that can help us determine if two variables should be (conditionally) independent under a specific causal DAG. We first define a *path* as a chain of non-repeating variables that is connected by arrows of either direction. For all paths between two variables:

1. If we are not conditioning on any variables, then a path is blocked if and only if somewhere in the path there are two arrowheads colliding. (We call the variable where the two arrowheads collide the *collider*)

2. If a path contains a non-collider that is conditioned, then the path is blocked.

3. If a collider or its descendent is conditioned, then the path is *not* blocked by the collider.

Based on the rules above, if there are any paths that are not blocked between the two variables of interest, then in the data they will be (conditionally) dependent. In epidemiology, we say that **confounding bias** is the bias introduced by unblocked paths formed by common causes between the treatment and the outcome, and **selection bias** is the bias introduced by stratifying on the collider and opening extra paths between the treatment and outcome.

6 Modeling in Causal Inference and Doubly Robust Estimators

In previous sections, we talked about the two approaches that can aid us in the estimation of ATE: standardization and inverse probability weighting. The building block of standardization is the conditional expectation of the outcome $\mu_a(l) = \mathbb{E}[Y|A = a, L = l]$, while for inverse probability weighting it is the conditional probability of treatment assignment $e_a(l) = \mathbb{P}[A = a|L = l]$. These two functions are often termed as *nuisance functions* since they are not directly of our interest. In the case where L has a small number of value patterns, both functions can be estimated with empirical means (which would be consistent as long as the conditional variance of Y is finite). However, in the case where L has a large number of value patterns, eg. some of the elements in L are continuous or L is high-dimensional, using empirical mean would be impractical or even infeasible, and we would need a model to aid us in the estimation of μ_a and e_a. Let's denote the estimators for μ_a and e_a as $\hat{\mu}_{a,n}$ and $\hat{e}_{a,n}$, where n is the number of observations used to estimate the nuisance functions.

Since we are using models to estimate μ_a and e_a, the "correctness" of the estimated models would be crucial. In particular, we would at least want them to be consistent. That is, for all $\varepsilon > 0$, $a \in \{0, 1\}$ and l in the support of L,

$$\lim_{n\to\infty} \mathbb{P}[|\hat{\mu}_{a,n}(l) - \mu_a(l)| > \varepsilon] = 0 \tag{54}$$

$$\lim_{n\to\infty} \mathbb{P}[|\hat{e}_{a,n}(l) - e_a(l)| > \varepsilon] = 0 \tag{55}$$

The consistency of $\hat{\mu}_{a,n}$ and $\hat{e}_{a,n}$ ensures that the plug-in estimators in Equations (42) and (51) are both asymptotically unbiased. When either of these models are mis-specified, their corresponding estimators for ATE will be asymptotically biased. It would thus be enticing to construct an estimator that is asymptotically unbiased as long as *one* of the nuisance function estimators is consistent. The **doubly robust** estimators serves this purpose exactly, and one of the doubly robust estimators is as follows

$$\hat{\tau}^{(dr)} = \frac{1}{N}\sum_i \hat{\tau}_i^{(dr)} = \frac{1}{N}\sum_i \Big[\hat{\mu}_1(L_i) - \hat{\mu}_0(L_i) + \Big(\frac{\mathbf{1}(A_i = 1)}{\hat{e}_1(L_i)} - \frac{\mathbf{1}(A_i = 0)}{\hat{e}_0(L_i)}\Big)(Y - \hat{\mu}_{A_i}(L_i))\Big] \tag{56}$$

To see why this estimator is doubly robust, first out of convenience, we assume that the nuisance functions are estimated in a separate, randomly split sample. Then, given $\hat{\mu}_a$ and $\hat{e}_a$, $\hat{\tau}^{(dr)}$ is an empirical mean on a sample independent to the sample used to derive the nuisance function. It is thus unbiased for

$$\mathbb{E}\Big[\hat{\mu}_1(L) - \hat{\mu}_0(L) + \Big(\frac{\mathbf{1}(A = 1)}{\hat{e}_1(L)} - \frac{\mathbf{1}(A = 0)}{\hat{e}_0(L)}\Big)(Y - \hat{\mu}_A(L))\Big] \tag{57}$$

where the expectation is taken over L, A, Y. Now for the doubly robust part, let's first assume that $\hat{\mu}_a$ is consistent and $\hat{e}_a$ is just a random function. Then, asymptotically we can replace $\hat{\mu}_a$ with μ_a, and the expectation becomes

$$\mathbb{E}\Big[\mu_1(L) - \mu_0(L) + \Big(\frac{\mathbf{1}(A = 1)}{\hat{e}_1(L)} - \frac{\mathbf{1}(A = 0)}{\hat{e}_0(L)}\Big)(Y - \mu_A(L))\Big] \tag{58}$$

$$= \mathbb{E}[\mu_1(L) - \mu_0(L)] + \mathbb{E}_L\Big[\mathbb{E}_{A,Y}\Big[\Big(\frac{\mathbf{1}(A = 1)}{\hat{e}_1(L)} - \frac{\mathbf{1}(A = 0)}{\hat{e}_0(L)}\Big)(Y - \mu_A(L))\Big|L\Big]\Big] \tag{59}$$

$$= \tau + \mathbb{E}_L\Big[\frac{\mathbb{E}_Y[Y - \mu_1(L)|A = 1, L]\mathbb{P}[A = 1|L]}{\hat{e}_1(L)}\Big] - \mathbb{E}_L\Big[\frac{\mathbb{E}_Y[Y - \mu_0(L)|A = 0, L]\mathbb{P}[A = 0|L]}{\hat{e}_0(L)}\Big] \tag{60}$$

$$= \tau + 0 - 0 = \tau \tag{61}$$

where in the last line we use the fact that $\mu_a(L) = \mathbb{E}[Y|A = a, L]$. Then let's assume that $\hat{e}_a$ is consistent and $\hat{\mu}_a$ is just a random function. Then asymptotically we can replace $\hat{e}_a$ with e_a, and

the expectation becomes

$$\mathbb{E}\Big[\hat{\mu}_1(L) - \hat{\mu}_0(L) + \Big(\frac{\mathbf{1}(A=1)}{e_1(L)} - \frac{\mathbf{1}(A=0)}{e_0(L)}\Big)(Y - \hat{\mu}_A(L))\Big] \tag{62}$$

$$= \mathbb{E}_L\Big[\mathbb{E}_{A,Y}\Big[\hat{\mu}_1(L) - \hat{\mu}_0(L) + \Big(\frac{\mathbf{1}(A=1)}{e_1(L)} - \frac{\mathbf{1}(A=0)}{e_0(L)}\Big)(Y - \hat{\mu}_A(L))\Big|L\Big]\Big] \tag{63}$$

$$= \mathbb{E}_L\Big[\hat{\mu}_1(L) - \hat{\mu}_0(L) + \frac{\mathbb{E}_Y[Y - \hat{\mu}_1(L)|A=1,L]\mathbb{P}[A=1|L]}{e_1(L)} - \frac{\mathbb{E}_Y[Y - \hat{\mu}_0(L)|A=0,L]\mathbb{P}[A=1|L]}{e_0(L)}\Big] \tag{64}$$

$$= \mathbb{E}_L\Big[\hat{\mu}_1(L) - \hat{\mu}_0(L) + \mathbb{E}_Y[Y - \hat{\mu}_1(L)|A=1,L] - \mathbb{E}_Y[Y - \hat{\mu}_0(L)|A=0,L]\Big] \tag{65}$$

$$= \mathbb{E}_L\Big[\hat{\mu}_1(L) - \hat{\mu}_0(L) + \mathbb{E}_Y[Y|A=1,L] - \hat{\mu}_1(L) - \mathbb{E}_Y[Y|A=0,L] + \hat{\mu}_0(L)\Big] \tag{66}$$

$$= \mathbb{E}_L[\mathbb{E}_Y[Y|A=1,L]] - \mathbb{E}_L[\mathbb{E}_Y[Y|A=0,L]] = \tau \tag{67}$$

where in the fourth line we use the fact that $e_a(L) = \mathbb{P}[A=a|L]$. In fact, it can be shown that, if we divide the doubly robust estimator into two parts, one estimating the mean counterfactual outcome under treatment 1 and another for treatment 0, then the asymptotic bias for the treatment a part is proportional to the *product* of the asymptotic biases for $\hat{\mu}_a$ and $\hat{e}_a$.

Another advantage of doubly robust estimators lies in the scenario where the dimension of L is large so that we require machine learning methods to estimate μ_a and e_a. When we use conventional statistical models without parameter regularization to estimate these functions, eg. linear regression, logistic regression, as long as the number of parameters does not grow relative to the sample size, the *mean squared error* of the function estimators are shrinking at an order of $n^{-1/2}$. Taking the sample mean as example: since it is unbiased for the population mean, its mean squared error is its standard deviation, $\sigma/\sqrt{n}$, where σ is the population standard deviation. Therefore, for standardization or IPTW estimators, if we use conventional statistical models to estimate the nuisance functions, the convergence rate for the estimators are in the order of $n^{-1/2}$. However, for modelling approaches that are more flexible like kernel smoothing or machine learning, the convergence rate of the function estimator, even under favorable condition (eg. sparsity of relevant variables, smoothness of the true function), the convergence rate would be something slower like $n^{-\delta}$ with $1/4 < \delta < 1/2$. This implies that standardization or IPTW with nuisance functions estimated by flexible approaches can have convergence rate slower than $n^{-1/2}$, leaving conventional robust standard errors and confidence intervals invalid. Since the doubly robust estimator involves products of the nuisance functions, it can be shown that as long as the *product* of convergence rate for the two nuisance functions is faster than $n^{-1/2}$ (which also implies that they should be consistent), their influence on the mean squared error for the resulting doubly robust estimator would be asymptotically negligible since the process of sample average would produce a mean squared error at the rate of $n^{-1/2}$.

An important side note is that, in order for the good convergence rate property to hold for doubly robust estimators, the data used to estimate the nuisance functions must be independent to the data used to calculate the sample average for ATE, like we have assumed in the previous derivation for doubly robustness. The easiest way to fulfill this is to do *sample-splitting*, where we use one random split of the data to estimate the nuisance functions and the remaining split to calculate the ATE. However, this would lead to waste of power since not all data are utilized to calculate the ATE. A better approach is to use cross-fitting, where the data is split into K folds. For each fold, the data other than that fold is used to estimate a set of nuisance functions, which is used only for that fold. Therefore, there will be K sets of nuisance functions, and in the calculation of ATE, the observations for each fold will use their own nuisance function.

To see how doubly robust estimators with cross-fitting work, let us see the following example:

```
library(AIPW)
library(SuperLearner)

## Loading required package:  nnls
## Loading required package:  gam
## Loading required package:  splines
## Loading required package:  foreach
## Loaded gam 1.22-2
```

```
## Super Learner
## Version:  2.0-28
## Package created on 2021-05-04

data("eager_sim_obs")

head(eager_sim_obs)

##   sim_A sim_Y eligibility loss_num   age time_try_pregnant      BMI    meanAP
## 1     0     0           0        1 21.96                 1 19.04475  81.33333
## 2     0     0           0        1 29.55                 6 35.56710 100.44444
## 3     1     1           1        2 26.34                 4 24.44728  84.00000
## 4     0     1           0        1 29.29                 3 25.50204  79.83333
## 5     1     0           1        1 25.68                 5 46.82850 105.66667
## 6     1     1           1        1 30.96                 0 23.35772  88.00000
```

Here we have a simulated dataset with `sim_Y` as the binary outcome of interest, `sim_A` as the binary treatment, and the other variables are the set of variables that serves as the candidates of confounders.

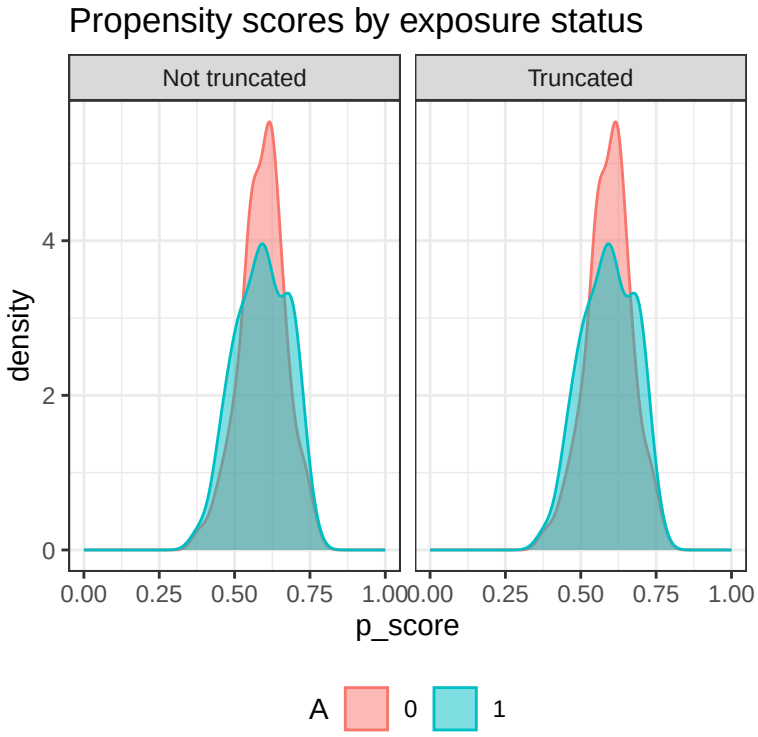
```
#Set the list of potential confounders
cov = c("loss_num", "age", "time_try_pregnant", "BMI", "meanAP")

#Set the SuperLearner libraries
sl.lib <- c("SL.xgboost", "SL.cforest")

#Construction of AIPW object
AIPW_SL <- AIPW$new(Y = eager_sim_obs$sim_Y,
                    A = eager_sim_obs$sim_A,
                    W = subset(eager_sim_obs, select=cov),
                    Q.SL.library = sl.lib,
                    g.SL.library = sl.lib,
                    k_split = 5,
                    verbose = T)

AIPW_SL$stratified_fit()

## Done!
```

Here we choose to use the `AIPW` package to carry out the doubly robust estimation, which by default calls the `SuperLearner` package to fit the two nuisance functions. Superlearner is a model fitting framework that uses cross-validation to automatically choose from different prediction algorithm. We specify two prediction algorithms for the package to choose from, `SL.xgboost` for XGBoost and `cforest` for random forest (there are numerous algorithms to choose from, see the manual for `SuperLearner`). `Y, A, W` specifies the outcome, treatment and list of variables to include in the two nuisance functions. If we would like different sets of covariates in $\hat{\mu}_a$ and $\hat{e}_a$, we may alternatively specify `W.Q` for the former and `W.g` for the latter. The `Q.SL.library` and `g.SL.library` specifies the list of algorithms used for the two nuisance functions. `k_split` is the number of splits for cross-fitting. The default is 10, but to save to we use 5. The `stratified_fit` method tells `AIPW` to fit $\hat{\mu}_1$ and $\hat{\mu}_0$ separately so that influence of the treatment variable is not regularized out.

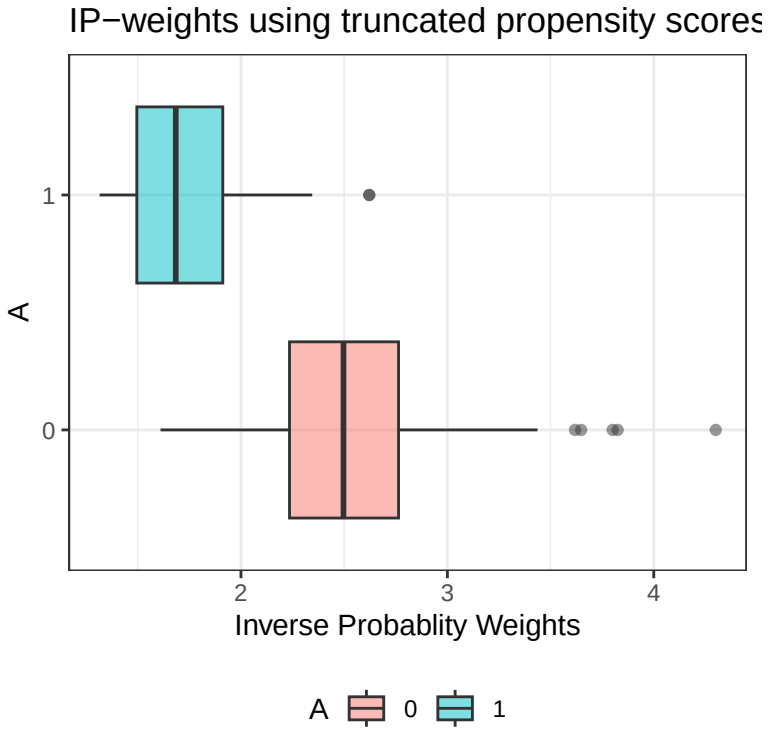
```
#Estimate the ATE, ATT and ATC
AIPW_SL$summary(g.bound = 0.025)$estimates #propensity score truncation

##                     Estimate     SE  95% LCL 95% UCL   N
## Risk of exposure      0.4414 0.0468  0.34962   0.533 118
## Risk of control       0.3302 0.0545  0.22348   0.437  82
## Risk Difference       0.1112 0.0698 -0.02567   0.248 200
## Risk Ratio            1.3368 0.1859  0.92850   1.925 200
## Odds Ratio            1.6029 0.2938  0.90116   2.851 200
## ATT Risk Difference   0.0911 0.0947 -0.09451   0.277 200
```

```
## ATC Risk Difference   0.1267 0.0627  0.00376   0.250 200
## $risk_A1
##   Estimate          SE    95% LCL    95% UCL
## 0.44143436 0.04684547 0.34961725 0.53325148
##
## $risk_A0
##   Estimate          SE    95% LCL    95% UCL
## 0.33022826 0.05446326 0.22348028 0.43697625
##
## $RD
##    Estimate           SE     95% LCL     95% UCL
##  0.11120610  0.06983641 -0.02567326  0.24808547
##
## $RR
##  Estimate        SE   95% LCL   95% UCL
## 1.3367553 0.1859330 0.9285014 1.9245147
##
## $OR
##  Estimate        SE   95% LCL   95% UCL
## 1.6028929 0.2938190 0.9011575 2.8510727
##
## $sigma_covar
##            [,1]       [,2]
## [1,] 0.59324929 0.02836199
## [2,] 0.02836199 0.43889953
```

```
library(ggplot2)
AIPW_SL$plot.p_score()
```




```
AIPW_SL$plot.ip_weights()
```



The `g.bound = 0.025` bit tells `AIPW` to truncate observations with $\hat{e}_1 < 0.025$ or $\hat{e}_1 > 1 - 0.025$. These observations, based on their covariate values, have a very high probability being assigned to one of the treatments, and it is pretty likely that population with these covariate values violates the positivity assumption. Therefore, we truncate them to avoid non-overlap issues, but the downside is that now we do not have a very intuitive explanation on which population we are inferring the ATE. Based on the results, the counterfactual event risk when everyone (except those truncated) is given the treatment is 44.19%, and when everyone is not given the treatment is 33.89%. Therefore, the ATE measured in risk difference is 10.30% with 95% confidence interval of $(-3.60\%, 2.42\%)$, which is not significant. The `ATT Risk Difference` is the *average treatment effect of the treated*, which is the average treatment effect within the population with `sim_A = 1`; the `ATC Risk Difference` is the *average treatment effect of the untreated (control)*, which is the average treatment effect within the population with `sim_A = 0`. This shows that there may *effect modification* for the actual treatment assignment, i.e. those received treatments are actually those benefiting less from the treatment. The first plot shows that distribution of propensity score has some overlap between the two treatment groups, which supports the positivity assumption, and the second plot shows that there are some participants not receiving the treatment with higher weights, but overall the weights are not extreme, which alleviates the concern of abnormally large weight influencing the results.