Fundamentals of Causal Inference with R: Module 1 Potential Outcomes Framework and Directed Acyclic Graphs

Babette Brumback

Professor Emerita
University of Florida, Biostatistics Department

Some Examples

- Throughout the short course, we refer to real and hypothetical data examples to motivate and illustrate the concepts and methods.
- We will describe some now.

Mortality Rates by Country

- Confounding is a term with many definitions. A google search provides a definition from the Oxford English dictionary (August 13, 2020) of 'confound' in standard English usage: "mix up (something) with something else so that the individual elements become difficult to distinguish".
- In statistics and related disciplines, a relationship between one variable and another was traditionally said to be 'confounded' when a third variable, often termed a 'lurking' variable, associated with both could explain part or all of the relationship, or could even reverse it.
- With the surge in popularity of causal inference, a more causal definition of confounding has become prevalent. Wikipedia's (August 13, 2020) first sentence under 'confounding' states: "In statistics, a confounder (also confounding variable, confounding factor, or lurking variable) is a variable that influences both the dependent variable and independent variable, causing a spurious association."
- In this course, we introduce what we will call a *true confounder*. We modify the Wikipedia definition somewhat, to define a true confounder as a variable that influences the exposure and that also influences the outcome via a directed path that does not include the exposure (presumably, it will also influence the outcome via a directed path that includes the exposure). This definition will be further explained later.

Mortality Rates by Country...

- The existence of one or more true confounders will typically cause *confounding*; that is, the *causal effect* of the exposure on the outcome cannot be identified without either (a) adjustment involving the true confounders, or, (b) adjustment involving other variables that we will refer to simply as confounders.
- In some cases, not all of the true confounders are required for identification of the causal effect. Confounders other than true confounders are only required when a *sufficient set of true confounders* is unavailable.

Mortality Rates by Country...

- One of the most basic examples often used to illustrate confounding is a comparison of mortality rates by countries. For example, consider the data in Table 1, which presents mortality rates by age group in the US and China in 2019 jointly with population figures from 2015.

Table 1: Mortality Rates by Age and Country

Country	Age Group	Deaths	Population	Rate
US	< 65	756,340	282,305,227	0.002679
US	65+	2,152,660	48,262,955	0.04460
US	Overall	2,909,000	330,568,182	0.0088
China	< 65	2,923,480	1,297,258,493	0.002254
China	65+	7,517,520	133,015,479	0.05652
China	Overall	10,441,000	1,430,273,973	0.0073

Mortality Rates by Country...

- The overall death rate in the US in 2019 is 8.8 per 1000, whereas that in China is only 7.3 per 1000. However, we can calculate that the percent of the US population aged 65 and over is 14.6%, compared with only 9.3% in China. Looking at the age-specific mortality rates shows that the older population in the US had a mortality rate of 4.46%, whereas that in China had a rate of 5.65%. The mortality rates for the younger populations are more similar, in absolute terms: 0.268% in the US versus 0.225% in China.
- This is nearly an example of *Simpson's paradox*, which refers to a relationship between two variables that reverses when a third variable is considered.
- That is, when we ignore age, the mortality rate is higher in the US than in China, but when we look within age category, the comparison reverses in the older age group (though not in the younger one).
- It is possible that if we were to break down the death rates further in the younger age group, we would find that they are uniformly lower in the US. In that case, the paradox would be complete.

National Center for Education Statistics 2018 Data

- The Integrated Postsecondary Education Data System (IPEDS) is a system of interrelated surveys conducted annually by the U.S. Department of Education's National Center for Education Statistics (NCES). IPEDS annually gathers information from about 6,400 colleges, universities, and technical and vocational institutions that participate in the federal student aid programs. Table 2 presents a subset of the admissions data from the NCES IPEDS access database provisionally for 2018-2019. The admissions data were collected in Fall of 2018; the table includes data from institutions that enrolled at least one man and woman in Fall 2018 and that also reported SAT scores.

Table 2: NCES Data

selective	female	highmathsat	*n*
0	0	0	435
0	0	1	87
0	1	0	420
0	1	1	37
1	0	0	50
1	0	1	104
1	1	0	55
1	1	1	29

NCES 2018...

- The variable `selective` indicates that the institution admitted less than 50% of applicants. An *indicator variable*, such as `selective`, for a condition is coded as 1 to indicate the condition and 0 to indicate the absence of the condition. The variable `female` indicates that more than 60% of students admitted were women. The variable `highmathsat` indicates that the average of the 25^{th} and 75^{th} percentiles of the math SAT scores for enrollees was greater than 600. The number of institutions represented by a row is n.

Reducing Alcohol Consumption: The What-If? Study

Table 3: The What-If? Study

T	A	H	Y	n
0	0	0	0	15
0	0	0	1	3
0	0	1	0	3
0	0	1	1	11
0	1	0	0	36
0	1	0	1	4
0	1	1	0	4
0	1	1	1	9
1	0	0	0	15
1	0	0	1	3
1	0	1	0	3
1	0	1	1	7
1	1	0	0	27
1	1	0	1	3
1	1	1	0	9
1	1	1	1	13

The What-If? Study...

- The WHAT-IF? (Will Having Alcohol Treatment Improve my Functioning?) study (Cook 2019) was a double-blind randomized clinical trial in which eligible women received either naltrexone 50 mg orally or placebo for 4 months, allocated in a 1:1 ratio, with assessments at baseline (i.e. at month zero, just before randomization) and at follow-up visits at 2, 4, and 7 months.
- Women living with HIV (WLWH) were eligible if they were 18 years or older and met past-month criteria for unhealthy alcohol use ($>$7 drinks/wk or $>$3 drinks on one single day at least twice).
- Naltrexone is an FDA-approved medication to help reduce drinking.
- In the dataset, $T = 1$ denotes naltrexone and $T = 0$ placebo, $A = 1$ denotes reduced drinking (less than or equal to 7 drinks per week) for the 30 days prior to month 4, $H = 1$ denotes unsuppressed HIV viral load ($\geq$ 200 copies/ml) at baseline, and $Y = 1$ denotes unsuppressed HIV viral load at month 4. The number of people represented in each row is given by n.
- This choice of notation will be used throughout the book, with T and A to denote observed or randomized treatments, H to denote the medical history influencing A or T, and Y to denote the outcome.

The What-If? Study...

- Questions of interest that can be addressed with What-If? study data are numerous, and we have presented just a subset of the available data.
- A primary aim of the study was to determine if naltrexone leads to reduced drinking at 4 months. This can be addressed with a chi-squared test (with the R function `chisq.test` specifying `correct=F`) of the association between T and A (65% for naltrexone versus 62.4% for placebo reduced drinking, P=0.72).
- One of the challenges with the What-If? study was that the participants were, in general, highly motivated to reduce drinking. Additionally, there was a small monetary incentive to participate, and participation required self report of heavy drinking, which meant that some participants may have consumed alcohol at a reduced rate even at baseline but reported it as heavy.
- These features implied limited chance of success for naltrexone to appear effective in this study population.
- One secondary aim was to determine if naltrexone induced improvements in clinical outcomes, such as viral load. In the naltrexone arm, 32.5% had unsuppressed viral load at month 4, versus 31.8% on placebo (P=0.92).
- Had this difference been larger, but still not statistically significant, we might have wished we had planned to use H, viral load at baseline, as a *precision variable* in our analysis, to reduce the variability in our comparison. Including a precision variable in the analysis can essentially subtract some of the variability in the outcome.

The What-If? Study...

- We can also investigate whether reducing drinking is associated with suppressed viral load; we observe 27.6% who reduced drinking have unsuppressed viral load at month 4, versus 40% who did not (P=0.1). A problem with this comparison is that the apparent effect of reducing drinking might be confounded.
- Once again, we will use H in the analysis, but this time as a confounder rather than as a precision variable.
- We will also use the What-If? study data to illustrate standardization and *doubly robust estimation*.
- Using just the binary data of Table 3, we will introduce *difference-in-differences estimators* and compare that approach to adjusting for confounding with standardization.
- We will also use the binary data to illustrate *instrumental variables estimation* as another alternative method for adjusting for confounding.

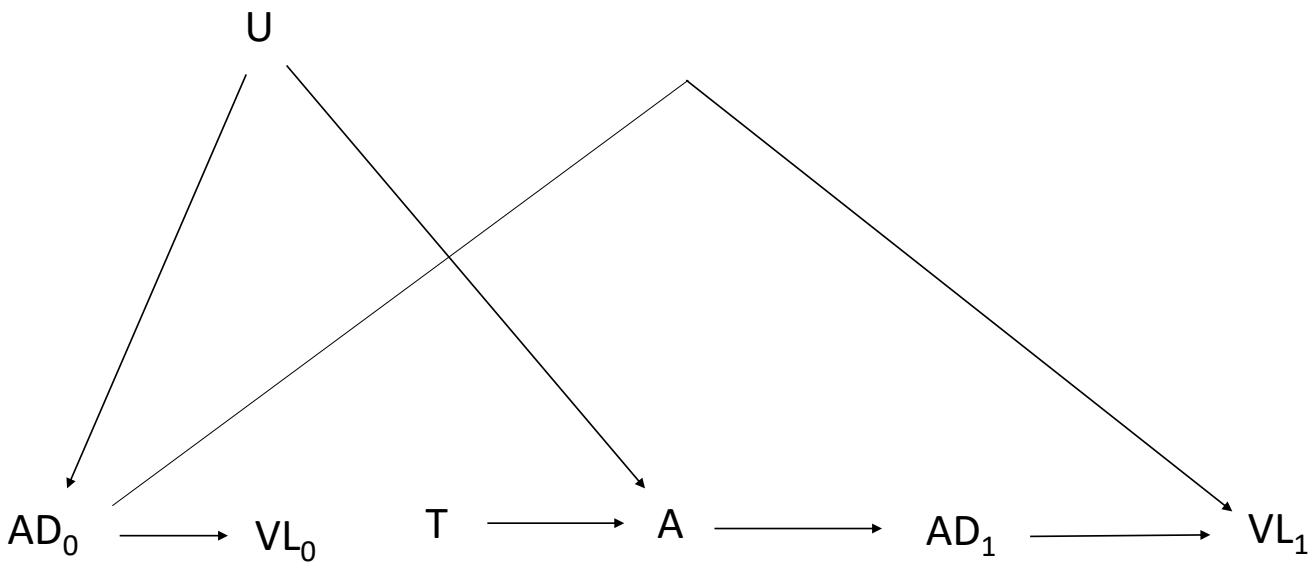
The Double What-If? Study

- It is not uncommon for human studies to go awry in one way or another. Humans are exceedingly complex and heterogeneous, and they vary over time in unpredictable ways. Indeed, one of the reasons to learn methods for causal inference is so that we can adjust for some of these vagaries in statistical analyses.
- However, it can be very helpful for learning and understanding the methods to actually *know* the causal mechanisms generating the data.
- For that reason, we introduce the **Double What-If? Study** (What If the What-If? Study had confirmed theory?) for illustration throughout the short course.

The Double What-If? Study...

- We simulated the data according to hyothetical known causal mechanisms depicted in Figure 1.

Figure 1: The Double What-If? Study



The Double What-If? Study...

- We will learn about *causal directed acyclic graphs* later, but an intuitive understanding of cause and effect represented via a single-headed arrow should suffice for an introduction to the study.
- For expository purporses, all variables are coded as binary.
- The variable U is latent, or unmeasured, variable representing a propensity for healthy behaviors ($U = 1$). The variables AD_0 and AD_1 represent adherence to HIV antiretroviral medication at times zero ($AD_0 = 1$) and one ($AD_1 = 1$). The variables VL_0 and VL_1 represent unsuppressed HIV viral load at times zero ($VL_0 = 1$) and one ($VL_1 = 1$).
- Note that antiretroviral adherence causes viral load even though it is coded at the same timepoint; adherence is summarizing a time period culminating at the timepoint whereas viral load is measured at that timepoint. The variable T represents randomization to naltrexone ($T = 1$) versus placebo ($T = 0$). There is no arrow into T because nothing causes it except for the randomization tool. The variable A represents reduced drinking ($A = 1$).
- We see that naltrexone causes reduced drinking, but so does a propensity for healthy behaviors, which also turns out to be associated with VL_1, viral load at time one (end of study) via a connection with AD_0, adherence to antiretrovirals at time zero.

The Double What-If? Study...

- Throughout the short course we will compare different methods for adjusting for confounding of the effect of A on VL_1, including standardization, difference-in-differences, and instrumental variables.
- We observe that the effect of A on VL_1 is *mediated* through AD_1. That is, the *mechanism* by which reduced drinking causes suppressed viral load is the influence of reduced drinking on adherence to antiretroviral medication, which in turn controls viral load.

The Double What-If? Study...

One of the beauties of investigating statistical methods with simulated data is that we know the truth. The R code for simulating the Double What-If? Study data is shown below.

```
doublewhatifsim.r<-function()
{
set.seed(444)
U<-rbinom(n=1000,size=1,prob=.5)
AD0prob<-.2+.6*U
AD0<-rbinom(n=1000,size=1,prob=AD0prob)
VL0prob<-.8 - .4*AD0
VL0<-rbinom(n=1000,size=1,prob=VL0prob)
T<-rbinom(n=1000,size=1,prob=.5)
Aprob<- .05+ T*U*.8
A<-rbinom(n=1000,size=1,prob=Aprob)
AD1prob<-.1+.8*A
AD1<-rbinom(n=1000,size=1,prob=AD1prob)
VL1prob<-VL0prob +.1 - .45*AD1
VL1<-rbinom(n=1000,size=1,prob=VL1prob)
dat<-cbind(AD0,VL0,T,A,AD1,VL1)
doublewhatifdat<-data.frame(dat)
doublewhatifdat
}
```

The Double What-If? Study...

- We set the randomization seed to 444 so that anyone running the code will produce the exact same data set.
- We simulated all binary variables using independent Bernoulli random variables with probability `prob`, where `prob` depends on the causal parents.
- The Bernoulli distribution models a weighted coin toss; independent Bernoulli random variables model repeated but unrelated weighted coin tosses.
- The Bernoulli distribution is a special case of the Binomial distribution, and therefore the `rbinom` with `size=1` generates Bernoulli random variables. We generate `n=1000` of them with probabilities given by `prob`, where the sample size of the study is `n=1000`.
- The variable names match the ones in Figure 1, except that when 'prob' is appended, it is not a random variable, but rather a collection of probabilities used for the weighted coin tosses for that variable. For example, we see that `VL0prob<-.8-.4*AD0`.
- This means that the probability of unsuppressed viral load is 0.8 without adherence to antiretrovirals at time zero, but with adherence, it is 0.8-0.4=0.4. Therefore, patients on antiretrovirals at time zero will be more likely to have suppressed viral load, or `VL0` equal to zero.

The RECOVERY Trial

- Coronavirus disease 2019 (COVID-19) causes diffuse lung damage that may progress to respiratory failure and death.
- The RECOVERY trial was a randomized, controlled, open-label, adaptive, platform trial comparing several possible treatments for COVID-19. We pulled results from a preliminary report on June 22, 2020 for the comparison of dexamethasone 6mg given once daily for up to ten days versus usual care alone in terms of 28-day mortality.
- Although the report presents a more sophisticated analysis, our simple comparison yields essentially identical results.
- The data are presented in Table 4 on the next slide, where $T = 1$ for dexamethasone and $T = 0$ for usual care, and $Y = 0$ for death within 28 days and $Y = 1$ for survival for at least 28 days; n is the number of patients per row.
- We calculated $\hat{E}(Y|T = 1) = 0.823$ and $\hat{E}(Y|T = 0) = 0.754$, with a chi-squared test of the difference yielding a p-value of 0.032. These results suggest that dexamethasone can increase 28-day survival, to some extent.

The RECOVERY Trial

Table 4: A Randomized Trial of Dexamethasone versus Usual Care

T	Y	n
0	0	1065
0	1	3256
1	0	454
1	1	2104

Potential Outcomes and the Consistency Assumption

- We will use the potential outcomes framework for causal inference in this short course.
- The RECOVERY trial raised a natural question for citizens of the world in the summer of 2020: if hospitalized for COVID-19, should we hope to receive dexamethasone? That is, would receiving it cause our survival?
- Thinking through this question carefully, what we really would like to do is compare our 28-day survival were we to receive dexamethasone with our 28-day survival were we to receive usual care.
- That is, we would like to compare our potential outcome $Y(1)$ to treatment with $T = 1$ with our potential outcome $Y(0)$ to treatment with $T = 0$.
- Throughout the text, we will use parenthetical notation, such as $Y(t)$, to denote potential outcomes, such as the potential outcome to $T = t$.
- We can contrast the potential outcomes $Y(0)$ and $Y(1)$ with the observable outcome Y that we would experience were we enrolled in the study.

Potential Outcomes and the Consistency Assumption

- The utility of the potential outcomes framework hinges on the validity of a *consistency assumption* that links potential outcomes to observed outcomes.
- This assumption states that if a participant is assigned $T = t$, the observed outcome Y equals, or is consistent with, the potential outcome $Y(t)$; that is, we can write $Y = Y(T)$.
- For a binary treatment, we can also write

$$Y = Y(0)(1 - T) + Y(1)T.$$

- Under the consistency assumption, we get to observe one potential outcome per study participant.
- While the potential outcomes are hypothetical and postulated to exist before treatment is assigned, the observed outcome does not exist until it is assessed, which is after treatment is assigned.

Potential Outcomes and the Consistency Assumption

- A major problem for causal inference is that we can only observe one potential outcome per participant.
- Therefore, direct comparisons like $Y(1)$ versus $Y(0)$ are impossible.
- Holland (1986) called this problem the *Fundamental Problem of Causal Inference* (FPCI).
- Thus, we can never know for sure whether dexamethasone would help us or not.
- An important feature of causal inference is that there must be a basis for comparison.
- Suppose we only observe $Y = Y(1) = 1$. Can we then say that dexamethasone caused that outcome?
- Without knowing $Y(0)$, we do not know whether the patient would have died without dexamethasone; if $Y(0) = Y(1)$, then receipt of dexamethasone does not matter, and it is not causal for that patient.
- In what follows, we will present two possible routes for circumventing the FPCI.

Potential Outcomes and the Consistency Assumption

- Recall that $Y = 0$ corresponds to death within 28 days and $Y = 1$ to 28-day survival.
- If we were able to know $Y(0)$ and $Y(1)$ for the patients in the study, we could classify them into four *causal types*:

$$\begin{aligned} \text{Doomed:} &\quad Y(0) = Y(1) = 0 \\ \text{Responsive:} &\quad Y(0) = 0, Y(1) = 1 \\ \text{Harmed:} &\quad Y(0) = 1, Y(1) = 0 \\ \text{Immune:} &\quad Y(0) = Y(1) = 1 \end{aligned}$$

- As we only observe Y, and not the potential outcomes, we cannot pinpoint the causal type.
- However, if a treated patient has $Y = 1$, the patient must be responsive or immune, whereas $Y = 0$, corresponds to doomed or harmed.
- For an untreated patient, $Y = 1$ implies harmed or immune, whereas $Y = 0$ implies doomed or responsive.
- For some studies, researchers may believe that the treatment cannot be harmful. In that case, a treated patient with $Y = 0$ must be doomed, while an untreated patient with $Y = 1$ must be immune.

Potential Outcomes and the Consistency Assumption

- It is worth emphasizing that the consistency assumption requires the potential outcomes to be well-defined.
- For example, in vaccine trials, even with a perfect vaccine, a participant's potential outcome to not getting the vaccine could depend on what percent of the study population is vaccinated and on how much that participant interacts with individuals outside of the study population.
- Let $Y = 1$ indicate infection and $Y = 0$ no infection, and $T = 1$ indicate receipt of a perfect vaccine and $T = 0$ no receipt.
- If everyone but one participant is vaccinated, and if that person only has contacts within the study population, then that person will not be exposed to the infectious disease, so that $Y(0) = 0$.
- On the other hand, if only half the study population is vaccinated, then that same unvaccinated participant will have a chance of exposure to the infectious disease, so that $Y(0)$ could equal 1.
- This example highlights that we cannot use the potential outcomes framework unless we can carefully define the potential outcomes.
- We could state that $Y(t)$ pertains specifically to a trial with half of the participants randomized to the vaccine. Extrapolation to a situation in which 80% of the population adopts a vaccine is tenuous.

Circumventing the Fundamental Problem of Causal Inference

- Holland (1986) presented two solutions to the fundamental problem of causal inference.
- The first he called the *scientific solution*, in which we could use scientific theory to measure both potential outcomes.
- For example, although we can only know just one of $Y_i(0)$ or $Y_i(1)$ for participant i of a study, if we believe participant j is essentially identical to participant i with respect to the two potential outcomes, we could assign $T_i = 1$ and $T_j = 0$ so that $Y_i = Y_i(1)$ and $Y_j = Y_j(0) = Y_i(0)$.
- Thus a contrast of the observables Y_i and Y_j would be a contrast of the unobservables $Y_i(1)$ and $Y_i(0)$.
- An example might be testing a new diet on genetically identical mice.
- Alternatively, an appeal to scientific theory might be used to impute $Y_i(0)$, the potential outcome to no treatment.
- For example, propofol ($T_i = 1$) is a commonly used sedative ($Y_i = 1$) for colonscopy; an anesthesiologist can safely assume that patient i would not be sedated with $T_i = 0$; that is, $Y_i(0) = 0$.
- After administering $T_i = 1$, we would observe $Y_i = Y_i(1)$ and thus whether propofol causes sedation.

Circumventing the Fundamental Problem of Causal Inference

- When we can assume that there are no *carry-over effects* of treatment T, we can observe individual i's outcome to $T_i = 0$ at one time point, i.e. $Y_{i1}(0)$ and to $T_i = 1$ at a second time point, i.e. $Y_{i2}(1)$, and assume either $Y_{i1}(0) = Y_{i2}(0)$ or $Y_{i2}(1) = Y_{i1}(1)$ so that a contrast such as $Y_{i2}(1) - Y_{i1}(0)$ essentially equals $Y_i(1) - Y_i(0)$.
- An example might be testing whether ibuprofen causes headache relief.

Circumventing the Fundamental Problem of Causal Inference

- Holland called the second solution the *statistical solution*, in which we randomly assign treatments to individuals.
- For a randomized clinical trial with two treatments (or a treatment and usual care or placebo) and an assignment probability that does not depend on any other data, randomization and the consistency assumption together imply that

$$(Y(0), Y(1)) \amalg T, \tag{1}$$

  because the potential outcomes exist prior to treatment, and the treatment is randomized independently of any existing data.
- The independence at (1) implies mean independence, that is, $E(Y(0)) = E(Y(0)|T = 0)$.
- Then, the consistency assumption implies that $E(Y(0)|T = 0) = E(Y|T = 0)$, so that $E(Y(0)) = E(Y|T = 0)$.
- Similarly, $E(Y(1)) = E(Y|T = 1)$.
- This means that we can estimate the expectation of the unobservables $E(Y(0))$ and $E(Y(1))$ with $\hat{E}(Y(0)) = \hat{E}(Y|T = 0)$ and $\hat{E}(Y(1)) = \hat{E}(Y|T = 1)$.
- Therefore, rather than contrasting the individual outcomes $Y(1)$ and $Y(0)$ for one person, we contrast the average outcomes $E(Y(1))$ and $E(Y(0))$ for the population.

Circumventing the Fundamental Problem of Causal Inference

- With the COVID-19 example, we can say that dexamethasone causes an increased chance of 28-day survival *on average* for the study population.
- Extrapolation of this result is fraught with difficulty.
- First, though the difference between 75.4% and 82.3% is statistically significant, it is still possible that it was due to chance.
- Second, the study population may no longer be relevant. New treatments might have been developed that were not in use at the time of study, or hospital protocols may have changed, or the population of patients may have changed; it is unknown how dexamethasone would work in conjunction with newer treatments or altered hospital protocols or a different case mix of patients.
- Third, and most important, extrapolating an average result to an individual is problematic.
- There may even be individuals in the study who were harmed by dexamethasone, but they were more than counterbalanced by individuals who were helped by it.
- And for the majority of patients, the outcome was the same with or without dexamethasone.

Circumventing the Fundamental Problem of Causal Inference

- For many scientific questions that cannot be addressed using the scientific solution to the FPCI, randomized studies are not feasible, due to ethical considerations or pragmatic reasons.
- In those cases, we are forced to rely on observational studies, and we must try our best to adjust for confounding.
- As we will see in the short course, there are several ways to do this, but a primary method relies on assuming that receipt of treatment effectively followed a *stratified randomized trial*, so that

$$(Y(0), Y(1)) \amalg A|H, \tag{2}$$

where $Y(a)$ is the potential outcome to $A = a$, A is completely randomized within strata defined by H, and H is the set of confounders thought to be important for adjustment.

- We also require a *positivity assumption*, that is,

$$1 > P(A = 1|H) > 0, \tag{3}$$

which asserts that individuals with a given level of H have a positive chance of receiving either $A = 1$ or $A = 0$.

Circumventing the Fundamental Problem of Causal Inference

- For example, suppose the obsertional study is prone to *confounding by indication*, that is, patients who are indicated for the treatment, and hence are sicker, tend to receive the treatment.
- In this case, the comparison group is healthier at the onset of the study.
- It should come as no surprise that treatments in such studies tend to appear harmful.
- However, if health prior to treatment is adequately captured by H, for example if $H = 0$ denotes unhealthy and $H = 1$ denotes healthy, then we might assume that, conditional on H, whether or not $A = 1$ is random.
- We might have that $P(A = 1|H = 0) = 0.7$ and $P(A = 1|H = 1) = 0.4$, but we can adjust for this in our analysis using the statistical solution to the FPCI.
- The conditional independence at (2) implies the mean independence $E(Y(0)|H) = E(Y(0)|H, A = 0)$, and the consistency assumption implies that $E(Y(0)|H, A = 0) = E(Y|H, A = 0)$, and thus we can estimate $E(Y(0)|H)$ with $\hat{E}(Y|H, A = 0)$.
- Note that the positivity assumption guarantees that in a large enough sample, there will be individuals in the $(H, A = 0)$ stratum to inform this estimate.
- Similarly, we can estimate $E(Y(1)|H)$ with $\hat{E}(Y|H, A = 1)$. Therefore we can contrast the conditional averages $E(Y(1)|H)$ with $E(Y(0)|H)$.

Circumventing the Fundamental Problem of Causal Inference

- We have shown how to use the conditional independence assumption at (2) to estimate conditional causal effects, such as $E(Y(1)|H)$ versus $E(Y(0)|H)$.
- Later, we will explain how to use it to estimate unconditional causal effects, such as $E(Y(1))$ versus $E(Y(0))$, and also a special kind of conditional causal effect, called the *average effect of treatment on the treated* (ATT), i.e. $E(Y(1)|A = 1)$ versus $E(Y(0)|A = 1)$.
- Note that if A were a completely randomized treatment as in (1), then the ATT would equal the overall treatment effect, $E(Y(1))$ versus $E(Y(0))$, because A would be independent of the potential outcomes.
- We will also use the method of differences in differences, the method of instrumental variables, and the front-door method to estimate the ATT.
- Finally, we note that one can also estimate the effect of treatment on the untreated using the same methods.

Effect Measures

- For binary Y, the most common effect measures are the *risk difference* (RD), *relative risk* (RR), and *odds ratio* (OR).
- Less common but also useful is what we will call the *other relative risk*, denoted by RR*, first brought to our attention by J.P. Scanlan.
- These four measures are defined as

$$\begin{aligned} \text{RD} &= p_1 - p_0 \\ \text{RR} &= \frac{p_1}{p_0} \\ \text{RR*} &= \frac{1-p_0}{1-p_1} \\ \text{OR} &= \frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0} \end{aligned}$$

- In the COVID-19 example, we estimated p_1 at 0.784 and p_0 at 0.754, therefore we have

$$\begin{aligned} \hat{\text{RD}} &= 0.069 \\ \hat{\text{RR}} &= 1.092 \\ \hat{\text{RR*}} &= 1.390 \\ \hat{\text{OR}} &= 1.293 \end{aligned}$$

Effect Measures

- These four measures always agree, qualitatively, in terms of the direction of the change in risk.
- In the COVID-19 example, dexamethasone increases the risk of 28-day survival from 0.754 to 0.784, and so the risk difference is positive, while the other three measures are greater than one.
- The other relative risk, RR*, deserves some explanation. Suppose the relative risk is greater than one, e.g. dexamethasone increases 28-day survival relative to withholding it.
- Then equivalently, the other relative risk must also be greater than one, e.g. withholding dexamethasone increases 28-day mortality relative to administering it.
- We can also use the RD, RR, RR*, and OR to measure association between two binary variables such as Y and T.
- We simply let $p_1 = E(Y|T = 1)$ and $p_0 = E(Y|T = 0)$.
- When the randomized clinical trial assumption (1) holds, the measures of association and the measures of causal effect are equal.
- When we need to distinguish the causal measures from the association measures, we will refer to the causal measures as causal RD, causal RR, causal RR*, and causal OR, whereas RD, RR, RR*, and OR will refer to the association measures.

Effect Measures

- Important aside: Whereas the effect measures will always agree qualitatively within a single population, or stratum (e.g. RD>0, RR>1, RR*>1, OR>1), one measure may present a treatment effect as stronger in a certain stratum, while another measure may present it as weaker in that same stratum.

Effect Measures

- There are other causal effect measures for binary outcomes that are important to consider, however, they can each be written as a function of just one of the RD, RR, or RR*.
- The *number needed to treat* (NNT) is

$$\frac{1}{p_1 - p_0} = \frac{1}{RD}.$$

  This measures the number of individuals we need to treat in order to benefit one.
- Suppose we treat a sample of size N. Then Np_1 is the number who we would expect to survive 28 days.
- Had we not treated that same sample, then we would expect only Np_0 to survive 28 days.
- The number needed to treat is the value for N that makes the difference $Np_1 - Np_0$ equal to one, i.e. $NNT(p_1 - p_0) = 1$, or $NNT = 1/(p_1 - p_0)$.
- We estimate that we would need to treat $1/0.069 = 15$ COVID-19 patients with dexamethasone in order to benefit one.
- Note that even though $1/0.069 = 14.49$, we did not round down to 14, because then NNT would be slightly less than one.

Effect Measures

- Three other causal effect measures that we will consider require us to assume that treatment cannot harm anyone; that is, there is no one of causal type 'harmed'.
- The *attributable fraction among the exposed*, which we will simply call the *attributable fraction* (AF), and which is also called the *probability of necessity* (PN), is
$$(p_1 - p_0)/p_1 = 1 - 1/RR.$$
- Suppose we treat a sample of size N, then, once again, Np_1 is the number we would expect to survive 28 days.
- Further, $Np_1 - Np_0$ is the number of those for whom treatment was necessary for survival.
- Therefore, the fraction of those who were treated and survived whose survival we could attribute to treatment (i.e. for whom treatment was necessary), is
$$N(p_1 - p_0)/Np_1 = (p_1 - p_0)/p_1.$$
- The probability of necessity is the conditional probability that treatment was necessary for suvival given treatment was administered and the individual survived.
- For the COVID-19 data, $PN = 0.084$.

Effect Measures

- The *causal power* (CP), also called the *probability of sufficiency* (PS), is

$$(p_1 - p_0)/(1 - p_0) = 1 - 1/\text{RR}^*.$$

- Suppose we withhold treatment from a sample of size N; then, $N(1 - p_0)$ is the number we would expect to die within 28 days.
- Further, $N(1 - p_0) - N(1 - p_1)$ is the number of those we would expect to have survived for 28 days had we treated them.
- Therefore, the fraction of those who were not treated and died within 28 days for whom treatment would have had the power to cause 28-day survival (i.e. for whom treatment would have been sufficient), is

$$\frac{N(1 - p_0) - N(1 - p_1)}{N(1 - p_0)} = \frac{p_1 - p_0}{1 - p_0}.$$

- The probability of sufficiency is the conditional probability that treatment would have been sufficient for survival given treatment was withheld and the individual died.
- For the COVID-19 data, $PS = 0.280$.

Effect Measures

- Finally, the *probability of necessity and sufficiency* (PNS) is

$$p_1 - p_0 = RD.$$

- We note that if we treat a sample of size N, then we expect $N(p_1 - p_0)$ to need treatment for survival, and $N(1 - p_0) - N(1 - p_1) = N(p_1 - p_0)$ is also the number for whom we would expect treatment is sufficient for preventing death.
- Therefore $N(p_1 - p_0)/N$ is the unconditional probability that treatment is both necessary and sufficient for survival (e.g. it is the proportion of 'responsive' individuals).
- For the COVID-19 data, $PNS = 0.069$.
- We note that PNS will always be less than both PN and PS.

Effect Measures

- The effect measures we have introduced to measure unconditional causal effects can also measure conditional causal effects; just repurpose p_1 and p_0 to denote conditional probabilities, such as $E(Y(1)|H = h)$ and $E(Y(0)|H = h)$.
- Parametric regression models are useful for estimating both unconditional and conditional causal effects.
- Indeed, the linear model has a natural correspondence with the risk difference, whereas the loglinear and logistic models have natural correspondences with the relative risk and odds ratio, respectively.
- For unconditional causal effects, suppose the randomized clinical trial assumption (1) holds so that $E(Y|T = 1) = E(Y(1))$ and $E(Y|T = 0) = E(Y(0))$.

Effect Measures

- The models are

$$\begin{aligned} \text{Linear} \quad & E(Y|T) = \beta_1 + \beta_2 T \\ \text{Loglinear} \quad & \log(E(Y|T)) = \beta_1 + \beta_2 T \\ \text{Logistic} \quad & \text{logit}(E(Y|T)) = \beta_1 + \beta_2 T \end{aligned} \tag{4}$$

- For the linear model, $\beta_2 = E(Y|T=1) - E(Y|T=0) = RD$.
- For the loglinear model, $\beta_2 = \log(E(Y|T=1)) - \log(E(Y|T=0)) = \log(RR)$, so that $\exp(\beta_2) = RR$.

Effect Measures

- For the logistic model, $\beta_2 = \text{logit}(E(Y|T=1)) - \text{logit}(E(Y|T=0))$, and $\text{logit}(p) = \log(o)$ where p is a risk and o is the odds $p/(1-p)$ corresponding to that risk.
- Therefore,

$$\beta_2 = \log \frac{E(Y|T=1)}{1-E(Y|T=1)} - \log \frac{E(Y|T=0)}{1-E(Y|T=0)} = \log(OR),$$

because

$$\log(o_1) - \log(o_0) = \log(o_1/o_0),$$

where o_1 and o_0 are two odds.

Effect Measures

- For conditional effect measures under assumption (2) with adjusting for $H = (H_1, H_2, H_3, H_4)$ and A replaced by T,
- the models are

$$\begin{aligned} \text{Linear} \quad & E(Y|T) = \beta_1 + \beta_2 T + \beta_3 H_1 + \cdots + \beta_6 H_4 \\ \text{Loglinear} \quad & \log(E(Y|T)) = \beta_1 + \beta_2 T + \beta_3 H_1 + \cdots + \beta_6 H_4 \\ \text{Logistic} \quad & \text{logit}(E(Y|T)) = \beta_1 + \beta_2 T + \beta_3 H_1 + \cdots + \beta_6 H_4 \end{aligned} \tag{5}$$

Effect Measures

- Because these are conditional measures, we need to specify values for $H_1, \ldots, H_4$.
- Some researchers set these equal to their averages, $E(H_1), \ldots, E(H_4)$, estimated from the data. However, for binary variables, this is awkward.
- We could, for example, set them all equal to one. Letting $H = (H_1, H_2, H_3, H_4)$, we denote $H = (1, 1, 1, 1)$ simply as $H = 1$.

Effect Measures

- Again, for the linear model,

$$\beta_2 = E(Y|T=1,H=1) - E(Y|T=0,H=1) = RD.$$

- For the loglinear model,

$$\beta_2 = \log(E(Y|T=1,H=1)) - \log(E(Y|T=0,H=1)) = \log(RR),$$

so that $\exp(\beta_2) = RR$.

- For the logistic model,

$$\beta_2 = \text{logit}(E(Y|T=1,H=1)) - \text{logit}(E(Y|T=0,H=1)),$$

so that

$$\beta_2 = \log \frac{E(Y|T=1,H=1)}{1-E(Y|T=1,H=1)} - \log \frac{E(Y|T=0,H=1)}{1-E(Y|T=0,H=1)} = \log(OR),$$

and $\exp(\beta_2) = OR$.

- These conditional effect measures provide one way to adjust for confounding by H, but in the next module we will use standardization, which yields an adjusted version of the unconditional effect measures.

Directed Acyclic Graphs: Theory

- Causal directed acyclic graphs provide a convenient and efficient way to represent statistical and causal dependence among a collection of variables.
- We rely heavily on Pearl (1995) for our overview.
- Given the collection $X_1, \ldots, X_p$, repeated application of the multiplication rule allows us to write their joint distribution as

$$P(X_1, \ldots, X_p) = \Pi_{j=1}^{p} P(X_j | X_1, \ldots, X_{j-1}), \tag{6}$$

  where X_0 is the empty set.
- The variables $(X_1, \ldots, X_{j-1})$ are termed the *predecessors* of X_j.
- Suppose $P(X_j | X_1, \ldots, X_{j-1}) = P(X_j | pa_j)$, where pa_j are a select group of the predecessors of X_j, termed the *parents* of X_j, and suppose also that pa_j is the minimal such set.
- For X_1, let pa_j be the empty set.
- Then we can write the joint distribution (6) as

$$P(X_1, \ldots, X_p) = \Pi_{j=1}^{p} P(X_j | pa_j). \tag{7}$$

- Notice that (7) encodes conditional independences not specified by (6). We can draw a directed acyclic graph (DAG) that also encodes these conditional independences.

- To do so, start with earlier variables and draw an arrow from X_h to X_j $(h < j)$ unless $X_j \amalg X_h | pa_j$.
- $$P(X_1, X_2, X_3, X_4) = P(X_4|X_2, X_1)P(X_3|X_2)P(X_2|X_1)P(X_1) \quad (8)$$
  corresponds to the DAG of Figure 2.

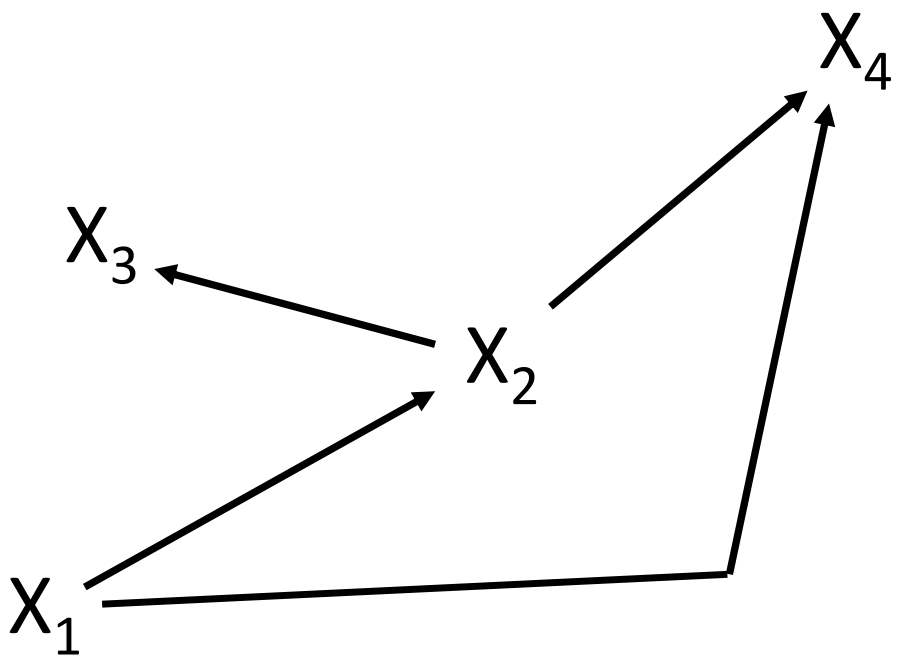


Figure 2: Example of a Directed Acyclic Graph

Theory

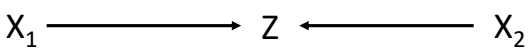
- We have constructed a graph encoding the conditional independences that are obvious from (8), which are $X_4 \amalg X_3 | X_2, X_1$ and $X_3 \amalg X_1 | X_2$.
- A useful theorem states that we can use the DAG to determine independences implied by the obvious ones that are not themselves obvious from (8).
- For our example, we also have that $X_4 \amalg X_3 | X_2$ and $X_3 \amalg X_1 | X_2, X_4$.
- These additional independences are difficult to determine directly from the factorization (8).
- One needs to apply a lot of algebraic manipulation. That is why the theorem allowing us to determine the conditional independences from the graphs is useful.

Theory

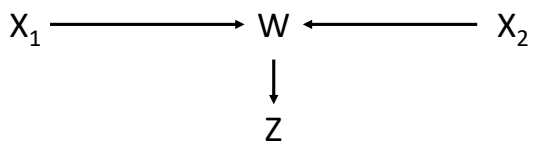
- To prepare us for the theorem, we need to learn *d-separation*.
- Consider 3 disjoint sets of variables A, B, and C, represented as nodes on a DAG.
- A *path* is a sequence of consecutive edges (of any directionality) in the graph.
- A path is said to be d-separated, or blocked, by a set of variables C if and only if the path (i) contains a *chain* as in Figure 3a, such that the middle variable Z is in C, or a *fork* as in Figure 3b, such that the middle variable Z is in C, or (ii) contains an inverted fork, or *collider*, as in Figure 3c such that the middle variable Z is not in C and such that no descendant of a collider is in C (i.e. in 3d, Z cannot be in C, and neither can be W.)
- Then the set C is said to *d-separate* A from B if and only if C blocks every path from a variable in A to a variable in B.

(a) Chain: Z is an Intermediate Variable

(b) Fork: Z is a Common Cause



(c) Collider: Z is a Common Effect



(d) Descendant of a Collider: Z is an Effect of a Common Effect

Figure 3: Graphical Structures

Theory

- Based on this definition of d-separation, the useful theorem can be stated as follows.

  Theorem: *If A and B are d-separated by C in a DAG, then* $A \amalg B|C$. *Conversely, if A and B are not d-separated by C in the DAG, then A and B are dependent conditional on C unless the dependences represented by the arrows exactly cancel.*

- The requirement that the dependences represented by the arrows do not exactly cancel is known as *faithfulness*.
- For help in applying this theorem, we will review the four basic graphical structures shown in Figure 3 and the independences implied by each.

Theory

- In Figure 3a, Z is an *intermediate variable*.
- The only independence implied by this structure is $X_1 \amalg X_2 | Z$.
- It is NOT true that $X_1 \amalg X_2$.
- In the context of the What-If? study, suppose X_1 indicates being randomized to treatment with naltrexone, Z indicates reducing drinking, and X_2 indicates subsequent health.
- It is reasonable to assume that naltrexone does not directly cause subsequent health, but indirectly can cause it via reduced drinking.
- Thus if our investigation is restricted to people who have reduced their drinking and have $Z = 1$, there would no longer be an association between having taken naltrexone and subsequent health; therefore, $X_1 \amalg X_2 | Z = 1$, and similarly for $Z = 0$.
- Because naltrexone can cause subsequent health indirectly, we have $X_1 \not\amalg X_2$.

Theory

- In Figure 3b, Z is a *common cause*.
- The only independence implied by this structure is $X_1 \amalg X_2 | Z$.
- It is NOT true that $X_1 \amalg X_2$.
- Suppose Z indicates healthy behaviors, which can cause adherence to HIV medication $X_1 = 1$ as well as healthy diet $X_2 = 1$.
- Supposing we know someone has adherence to HIV medication, we would predict he or she had healthy behaviors, and therefore we would also predict healthy diet; thus, $X_1 \not\amalg X_2$.
- However, suppose we look only at people with healthy behaviors $Z = 1$; then we would expect everyone to be adherent to HIV medication and have a healthy diet.
- With X_1 and X_2 constant, there is no statistical association; that is, knowledge of adherence to HIV medication does not effect prediction of healthy diet, because we know that everyone has a healthy diet.
- Therefore, $X_1 \amalg X_2 | Z = 1$.
- Analogously, for $Z = 0$, we would expect everyone to fail to adhere to their medication and not have a healthy diet, so that $X_1 \amalg X_2 | Z = 0$.

Theory

- In Figure 3c, Z is a *common effect*.
- The only independence implied by this structure is $X_1 \amalg X_2$.
- It is NOT true that $X_1 \amalg X_2 | Z$.
- Suppose X_1 indicates a selective university and X_2 indicates admitting proportionally more men, and Z indicates a high average math SAT score of enrollees.
- Knowing a university is selective does not give us information on whether or not it admits proportionally more men, so that $X_1 \amalg X_2$.
- However, If we look only at universities with high average math SAT scores, knowing that the university is not selective would induce us to predict that it admits proportionally more men, so that $X_1 \not\amalg X_2 | Z$.

Theory

- In Figure 3d, Z is an *effect of a common effect*.
- The independences implied by this structure are $X_1 \amalg X_2$, $Z \amalg X_1 | W$, and $Z \amalg X_2 | W$.
- It is NOT true that $X_1 \amalg X_2 | W$ or that $X_1 \amalg X_2 | Z$.
- This is the trickiest of the four types of structures.
- Now let X_1 indiciate a selective university, X_2 indicate admitting proportionally more men, W indicate high average math SAT score of enrollees, and Z indicate high average math GRE score of graduates.
- $X_1 \amalg X_2$ for the same reasons as in the previous example, but $X_1 \not\amalg X_2 | Z$ because knowing the school has high average math GRE scores of graduates induces us to predict it has high math SAT scores of enrollees, and therefore X_1 and X_2 are associated again given Z for the same reasons as in the previous example.

Theory: Practice Applying the Theorem

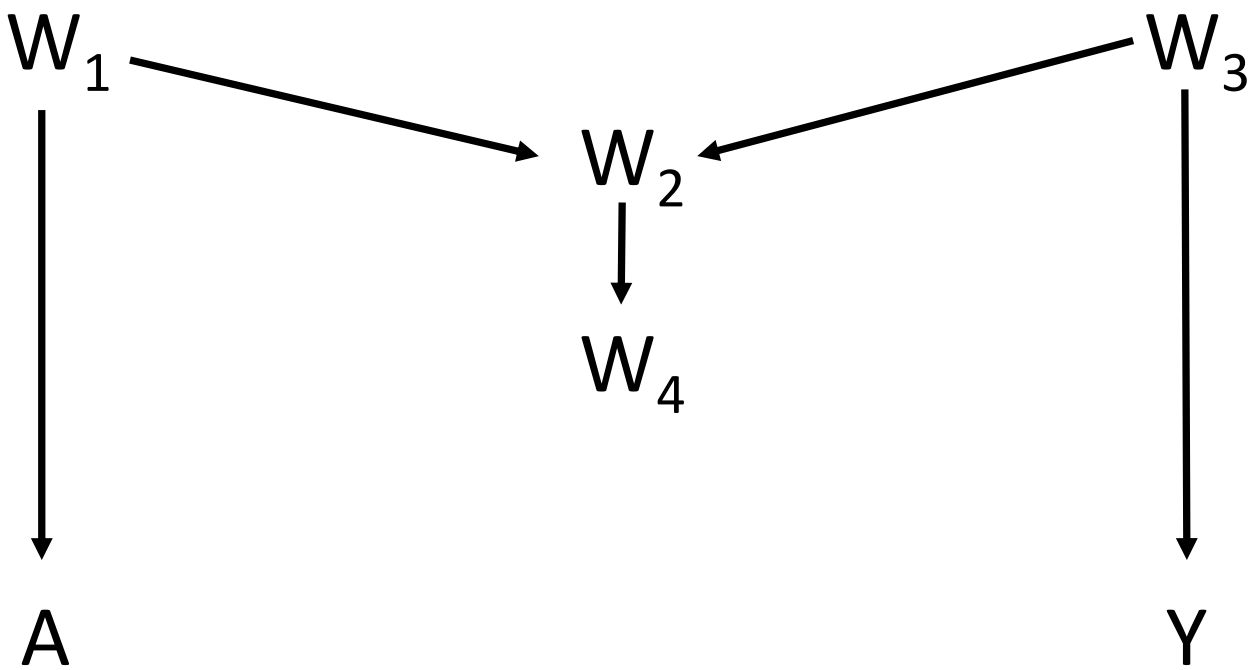


Figure 4: Causal DAG With Collider

Theory: Practice Applying the Theorem

- To practice applying the theorem, consider the DAG of Figure 4. Some of the independences that the DAG implies are

$$
\begin{aligned}
A &\perp\!\!\!\perp Y \\
A &\perp\!\!\!\perp W_3 \\
A &\perp\!\!\!\perp Y \mid W_3 \\
A &\perp\!\!\!\perp W_3 \mid Y \\
A &\perp\!\!\!\perp Y \mid W_1, W_2 \\
A &\perp\!\!\!\perp Y \mid W_4, W_3 \\
W_1 &\perp\!\!\!\perp W_3 \\
W_1 &\perp\!\!\!\perp Y \\
W_1 &\perp\!\!\!\perp Y \mid A \\
W_4 &\perp\!\!\!\perp Y \mid W_2.
\end{aligned}
$$

Theory: Practice Applying the Theorem

- Some of the dependences that the DAG implies, assuming faithfulness, are

$$
\begin{aligned}
A &\not\perp\!\!\!\perp W_1 \\
A &\not\perp\!\!\!\perp W_2 \\
A &\not\perp\!\!\!\perp Y \mid W_2 \\
A &\not\perp\!\!\!\perp Y \mid W_4 \\
A &\not\perp\!\!\!\perp W_1 \mid W_2, W_4 \\
W_1 &\not\perp\!\!\!\perp W_2 \\
W_1 &\not\perp\!\!\!\perp W_4 \\
W_1 &\not\perp\!\!\!\perp W_3 \mid W_2 \\
W_1 &\not\perp\!\!\!\perp W_3 \mid W_2, W_4 \\
W_3 &\not\perp\!\!\!\perp Y.
\end{aligned}
$$

Theory: Practice Applying the Theorem

- One might wonder if the DAG of Figure 4 could correspond to any real example.
- Indeed, it is quite difficult to come up with real examples in which the missing arrows are plausible.
- Consider the following attempt to characterize a randomized trial in a population of post-menopausal women having difficulty sleeping.
- Let W_1 be randomization to a certain sedative, which sometimes causes headache disorder, A, but can induce more than 6 hours of sleep per night, W_2.
- Let W_3 be randomization to HRT supplementation, which can also aid in sleep (W_2) due to reduction of hot flashes, but carries a risk of increased blood pressure, Y.
- Finally, let W_4 indicate increased productivity at work.
- One might wonder if the headache disorder A and the increased blood pressure Y are related; the DAG says they are not.
- The difficulty with this example, as with many others, in satifying the assumptions of the DAG, is that sleeping less than 6 hours per night might induce a headache disorder, and a headache disorder might reduce productivity at work, and less sleep as well as unproductive work days might increase blood pressure.
- It is often too hard to justify the missing arrows.

Theory

- The previous theorem is most useful because it implies another theorem which helps us to identify causal effects by identifying a sufficient set of confounders.
- Specifically,

  Backdoor Theorem: *Given a DAG containing two variables A and Y as well as a set of variables C excluding A and Y that does not contain descendants of A nor Y, the set C is sufficient to adjust for confounding of the effect of A on Y if and only if there exists* no unblocked backdoor path *from A to Y. That is,*

  $$\{Y(a)\}_{a \in \mathcal{A}} \amalg A | C,$$

  where $\{Y(a)\}$ is the set of potential outcomes to all possible values $a \in \mathcal{A}$ of A.

Theory

- Again considering our example DAG of Figure 4, the empty set (that is, no variables in C) is sufficient to adjust for confounding of the effect of A on Y, but so is the set W_1 or the set (W_1, W_2) or (W_3, W_4).
- It is noteworthy that the variable W_2 is insufficient by itself to adjust for confounding of the effect of A on Y.
- This insufficiency proves that a traditional definition of confounder as a variable that is associated with A and Y is inadequate, because W_2 is such a variable and because the DAG implies that adjusting for it is worse than not adjusting for any variables.
- Adjusting for W_2 or W_4 without also adjusting for W_1 or W_3 can cause a bias referred to as *collider bias*.

Theory

- Referring back to our definition of *true confounder*, i.e. a variable that influences the exposure and that also influences the outcome via a directed path that does not include the exposure, we see that W_2 does not satisfy the criteria, as it influences neither A nor Y.
- In fact, none of the variables in the DAG of Figure 4 satisfies the criteria, and therefore, there is no true confounder and hence no confounding of the effect of A on Y.
- We learn from this example that a sufficient set of confounders need not contain a true confounder.
- However, if there is no true confounder in a DAG, then the empty set is a sufficient set of confounders.
- As mentioned previously, in some cases not all of the true confounders are required to form a *sufficient set of true confounders*, which is simply a sufficient set of confounders that are all true confounders.

Examples

- We conclude with the causal DAG used to generate data for the Double What-If? Study, introduced previously.
- Several questions could be answered using data from the study.
- For example, one might ask what is the *intent-to-treat* effect of naltrexone ($T = 1$) on end of study viral load (VL_1)?
- The intent-to-treat effect, discussed at more length later, simply compares the outcome across the two randomized groups, $T = 1$ and $T = 0$, without concern whether participants assigned to naltrexone actually took naltrexone.
- The causal DAG shows that naltrexone was randomized, as there are no arrows into T.
- Therefore, we do not need to worry about confounding of the effect of T on VL_1.
- However, if we were interested in the effect of reducing drinking ($A = 1$) on VL_1, we need to worry about the true confounder U, which has an arrow into A and a directed path from U to VL_1 mediated by AD_0.

Examples

- The Backdoor Theorem tells us that AD_0 is sufficient to adjust for confounding of the effect of A on VL_1; letting $VL_1(0)$ and $VL_1(1)$ be the potential outcomes to $A = 0$ and $A = 1$, the Backdoor Theorem tells us that $\{VL_1(0), VL_1(1)\} \amalg A | AD_0$.
- The Backdoor Theorem also tells us that VL_0 is insufficient to adjust for that confounding; that is, $\{VL_1(0), VL_1(1)\} \not\amalg A | VL_0$.
- Soon, we will learn how to use standardization to adjust for the effect of A on VL_1.
- Later, we will compare standardization to a difference-in-differences approach to adjust for confounding, which makes use of VL_0 rather than AD_0, and is valid under a different assumption.
- We introduce three simple difference-in-difference approaches to adjusting for confounding, and the one that is valid for the Double What-If? Study relies on the assumption that

$$E(VL_1(0)|A = 1) - E(VL_1(0)|A = 0) = E(VL_0|A = 1) - E(VL_0|A = 0),$$

  called *additive equiconfounding*.
- Thus, the Backdoor Theorem gives us one way to adjust for confounding, but there are other ways to consider as well.
- Yet another method uses T as an *instrumental variable* to adjust for confounding of the effect of A on VL_1.

Causal DAGs versus Associational DAGs

- In the preceding slides we have been implicitly assuming that the Double What-If simulation and the DAGs under discussion were causal.
- However, they could just as equally be associational.
- To illustrate this, we consider the simplest DAG with confounding of the effect of A on Y by H and illustrate how placement of the potential outcomes $(Y(0), Y(1))$ on the graph can toggle it from causal to associational.

Causal DAGs versus Associational DAGs

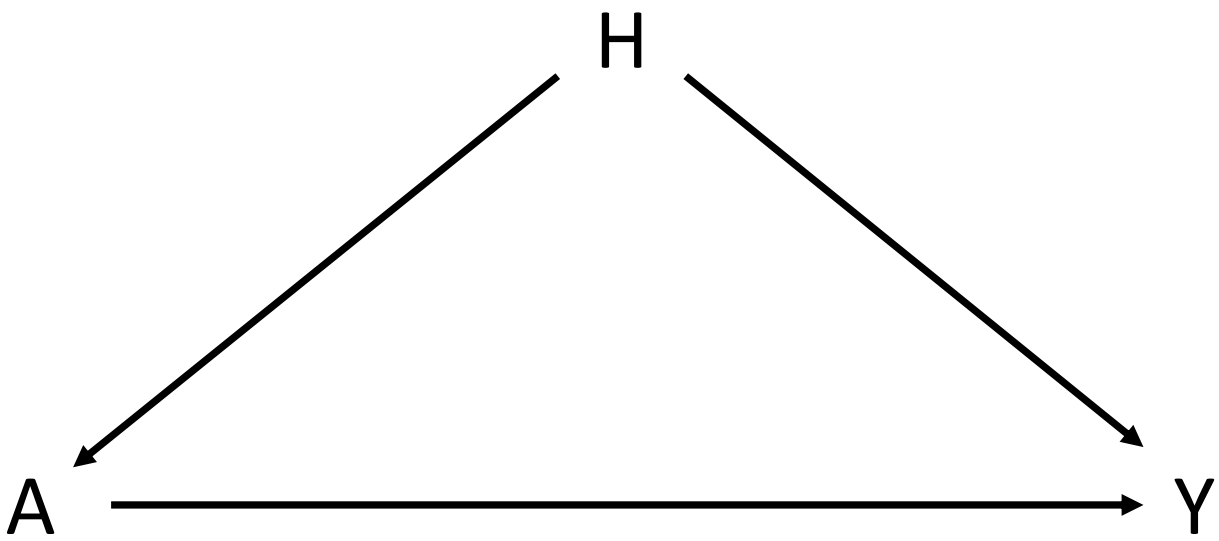


Figure 5: Associational or Causal DAG?

Causal DAGs versus Associational DAGs

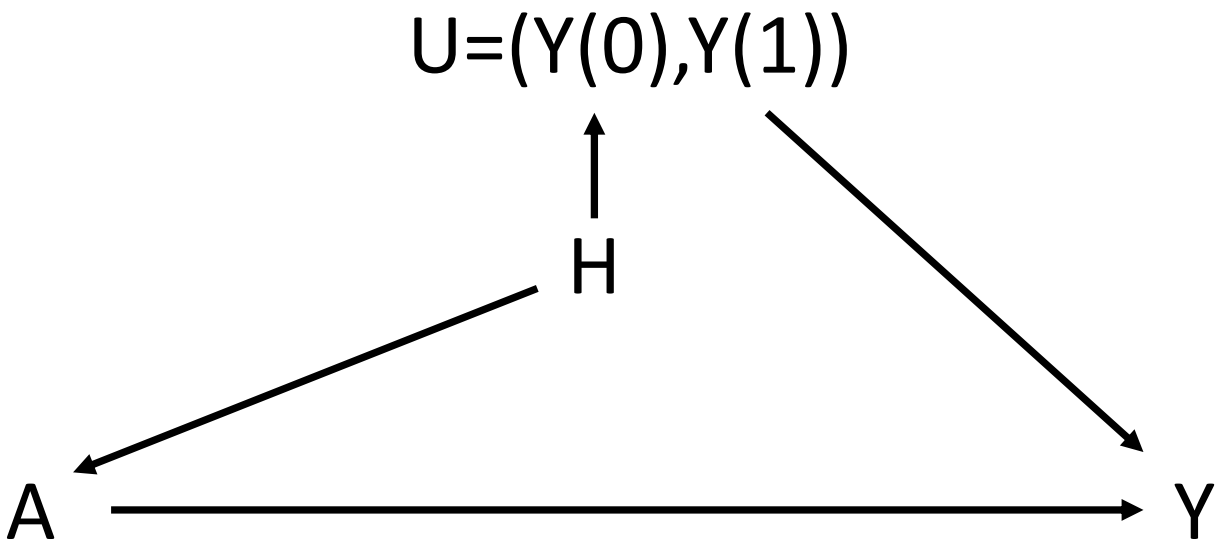


Figure 6: The Previous DAG was Indeed Causal

Causal DAGs versus Associational DAGs

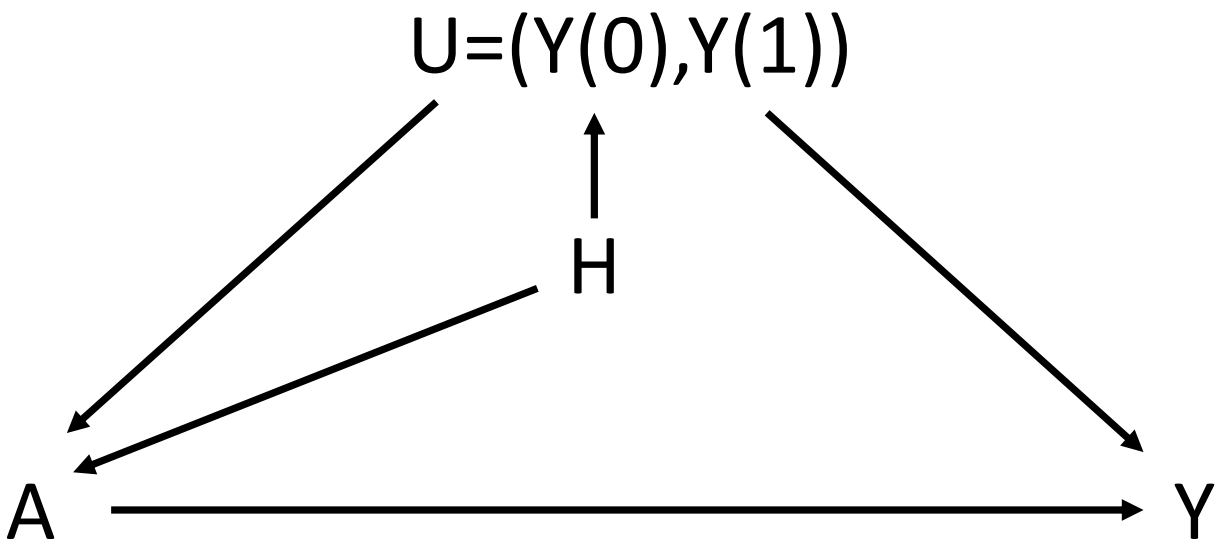


Figure 7: The Previous DAG had to have been Associational

Conclusions

- Although a correct causal DAG is very useful for mapping out an appropriate statistical analysis plan, correctly constructing a causal DAG for a scientific study is generally quite difficult.
- In most cases, the causal DAG is likely to be misspecified.
- For example, in Figure 4, an arrow from W_2 or W_4 to A or Y would change the confounding structure drammatically.
- Rather than having no confounding, there would be a true confounder and we would need to adjust our analysis.
- Supposing the arrow were from W_4 to Y, then W_1 would be a true confounder and also sufficient for confounding adjustment.
- Supposing the arrow were from W_4 to A, then W_3 would be a true confounder and also sufficient for confounding adjustment.
- Supposing arrows both from W_4 to A and from W_4 to Y, then (W_1, W_2, W_3, W_4) would be a true confounder and also sufficient for confounding adjustment.
- It may be the case that causal DAGs are more useful for destructive purposes than constructive purposes.
- Given a putative causal DAG and corresponding statistical analysis plan, a reviewer can argue that missing arrows should instead be present, or that other variables need to be inserted into the DAG.
- The reviewer's DAG will often show the original statistical analysis plan to be biased.
- However, even if the impact is to induce caution rather than to offer validation, causal DAGs can play an important role in causal inference.

Hill 1965

- Due to all of the assumptions required for proper causal inference, some common-sense guidelines are also important to bear in mind. In 1965, the statistician and epidemiologist Sir Austin Bradford Hill provided nine criteria to assist in making causal judgements from observational data.
- Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?
- (1) Strength. First upon my list I would put the strength of the association.
- (2) Consistency: Next on my list of features to be specially considered I would place the consistency of the observed association. Has it been repeatedly observed by different persons, in different places, circumstances and times?

Hill 1965...

- (3) Specificity: One reason, needless to say, is the specificity of the association, the third characteristic which invariably we must consider.
- (4) Temporality: My fourth characteristic is the temporal relationship of the association –which is the cart and which the horse?
- (5) Biological gradient: Fifthly, if the association is one which can reveal a biological gradient, or dose-response curve, then we should look most carefully for such evidence.
- (6) Plausibility: It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand. What is biologically plausible depends upon the biological knowledge of the day.
- (7) Coherence: On the other hand the cause-and-effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease - in the expression of the Advisory Committee to the Surgeon-General it should have coherence.

Hill 1965...

- (8) Experiment: Occasionally it is possible to appeal to experimental, or semi-experimental, evidence. For example, because of an observed association some preventive action is taken. Does it in fact prevent?
- (9) Analogy: In some circumstances it would be fair to judge by analogy. With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy