Formal Verification of the Backdoor Adjustment Formula

by

Evgeniya Artemova

S.B. Mathematics and Electrical Engineering and Computer Science, MIT, 2024.

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2026

Authored by: Evgeniya Artemova
Department of Electrical Engineering and Computer Science
May 8, 2026

Certified by: Adam Chlipala
Arthur J. Conner Professor of Computer Science, Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Formal Verification of the Backdoor Adjustment Formula

by

Evgeniya Artemova

Submitted to the Department of Electrical Engineering and Computer Science
on May 8, 2026 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Understanding cause-and-effect relationships is a difficult problem and one that is central to modern-day science. One such cause-and-effect relationship that we are often concerned with is the effect of interventional measures, such as government policies, where we force some choice or decision onto the population. Often the effects of interventions need to be predicted from observational data, since an experiment that would more directly acquire the desired data, such as a randomized controlled trial, may be unethical or impractical to perform. In these cases, one must consider other factors and influences that affect the observational data. One method to record and reason about relationships between variables is to use structural causal models. Many theorems have been proven about structural causal models. In this thesis we focus on one such theorem, the backdoor adjustment formula, which gives us a way to predict the outcomes of interventional measures. We use Rocq to verify the backdoor adjustment formula for certain structural causal models as well as present the beginning of a formally verified proof of the backdoor adjustment formula in the general case. We also show the applicability of the backdoor adjustment formula to science experiments by discussing some papers and proving the formula they could have used to predict the effects of an intervention.

Thesis supervisor: Adam Chlipala
Title: Arthur J. Conner Professor of Computer Science

Acknowledgments

I am extremely grateful to my academic advisor, Adam Chlipala, for introducing me to both formal verification and causal reasoning. I greatly enjoyed learning more about these fields through this research project. His guidance and support was crucial both to making this research possible and ensuring I had a fun time doing it.

I would also like to thank the rest of our research group, including Julia Turcotti and Blanca Luo at MIT, Eunice Jun and London Bielicke at UCLA, and Emery Berger at UMass Amherst. Thank you for all of the input you have given to this project and all of the guidance you have provided.

I am also grateful to Brynmor Chapman and the rest of the 6.1210 staff for giving me the chance to be a teaching assistant for the last four semesters. This opportunity has provided funding for this research, as well as given me lots of rewarding and memorable experiences.

I would like to express my gratitude to my friends, with whom I experienced the highs and lows of MIT. Thank you to Tetazoo, my undergraduate living community, for the adventures it provided me with. I would also like to thank the handful of friends with whom I spent many hours coworking over the past year, and who were instrumental in making this thesis exist, including Rory, Kiran, and Kat.

Finally, thank you to my family for their steady support and love. I would especially like to thank my parents and uncle, who were excited about and happy to discuss my research with me.

Some of the Rocq proofs in this thesis, especially lemmas involving simple mathematical statements, were completed with the assistance of LLMs.

Contents

List of Figures 9

1 Introduction **11**
1.1 Formal Verification . . . 12
1.2 Background and Related Work . . . 12
1.2.1 Structural Causal Models . . . 13
1.2.2 Interventions . . . 14
1.2.3 Backdoor Adjustment Formula . . . 15
1.2.4 Proofs of the Backdoor Adjustment Formula . . . 18
1.2.5 Existing Formalizations of Experiment Design . . . 19

2 Backdoor Adjustment Formula for Example Graphs **21**
2.1 Node Functions and the Do-Operator in Rocq . . . 21
2.2 Two-Node Model . . . 22
2.3 Mediator . . . 25
2.4 Confounder . . . 27
2.5 Collider . . . 29
2.6 Implementation Details . . . 30

3 Backdoor Adjustment Formula in the General Case **31**
3.1 Joint Probability Distributions and Graph Factorization . . . 31
3.2 Parental Adjustment Formula . . . 34
3.3 Backdoor Adjustment Formula from Parental Adjustment Formula . . . 38
3.4 Example Graphs . . . 42
3.4.1 Four-Node Example Graph . . . 42
3.4.2 Eight-Node Example Graph . . . 44
3.5 Implementation Details . . . 45

4 Application to Scientific Experiments **47**
4.1 Electronic Media and GPA . . . 47
4.1.1 Paper Results . . . 47
4.1.2 Causal Interpretation . . . 48
4.2 Alcohol Advertisement and Alcohol Consumption . . . 49
4.2.1 Paper Results . . . 50
4.2.2 Causal Interpretation . . . 50

5 Future Work **53**
5.1 Backdoor Adjustment Formula Proof 53
5.2 Exploring Assumptions in Theorems 54
5.3 Connecting to the Semantic Meaning of a Structural Causal Model 54
5.4 Interventional and Observational Relationships Beyond the Backdoor Adjustment Formula 54
5.5 Formally Verifying Experimental-Design Validity 55

References 57

List of Figures

1.1 A sample structural causal model showing the effects 13
1.2 A mediator, confounder and collider 16
1.3 An example structural causal model 17

2.1 A two-node structural causal model 22
2.2 A three-node structural causal model with a mediator 26
2.3 A three-node structural causal model with a confounder 27
2.4 A three-node structural causal model with a collider 29

3.1 Two structural causal models, one with and one without intervention 32
3.2 Backdoor paths between T and H, all of which include parents 35
3.3 Options for backdoor paths between T and Z 41
3.4 Options for backdoor paths between H and $\mathrm{PA}(T)$ 42
3.5 A four-node structural causal model with two confounders 43
3.6 An eight-node structural causal model with sets Z, $\mathrm{PA}(T)$ and E of size greater than one 44

4.1 Causal model showing variables that affect electronic media use and GPA, as well as causal relationships between them 48
4.2 Causal model showing the effect of alcohol advertisements on alcohol consumption 51

Chapter 1

Introduction

Experiments are a fundamental tool we use for understanding the world and the cause-and-effect relationships that govern it. For any experiment, someone must design it, developing a design that will showcase the desired causal relationship, while attempting to remove as much bias as possible. Despite the prevalence of experiments, there is not as much standardization in experiment design as one might expect. The lack of standardization leads to many issues with experimental validity and reproducibility, where one finds that the results of an experiment cannot be recreated or are in some other way dubious. In fact, it is not uncommon that different experiments that aim to study the same effect find seemingly contradictory results.

One such example is hormone replacement therapy (HRT) in post-menopausal women. Initially, some studies found positive outcomes for HRT in women and supported the use of this treatment. Later, a randomized controlled trial was conducted that found that there was a higher rate of death for women on HRT, which resulted in the abandonment of HRT for post-menopausal women [1]. Since these experiments yielded different results, it is clear that the conclusion drawn from at least one of them was not entirely correct. A common cause for such inconsistencies is not explicitly noting all the assumptions being made. It may be hard to account for all of the variables that might influence the outcome, and when such assumptions are not explicit, results may appear contradictory even when they are not.

When these two HRT experiments were reanalyzed, it was found that there were ways to reconcile these results as long as one paid attention to the differences in setup and explicitly stated them in the final results [2]. Specifically, some of the contradictions came from the fact that one experiment observed health risks in the immediate aftermath of starting HRT, while the other observed them after subjects had been undergoing HRT for some period of time. Also, there were other factors, such as the time since the onset of menopause, that were influencing results but were not accounted for in the analysis. While these factors were important-enough to cause the experiments to yield different results, they were subtle-enough that they were initially not remarked upon.

In this thesis, we consider experiments through the lens of structural causal models, since this approach lets us write out, and work with, all of our assumptions about confounding. Structural causal models require the explicit statement of all the variables being considered and relationships between them, which makes it easier to resolve apparent inconsistencies, as in the experiment above. We focus specifically on the problem of how one can predict the outcomes of *interventional* measures, such as government policies, in the case where

one cannot use a randomized controlled trial and must use *observational* data to reach a conclusion. We work on proving a specific theorem, the *backdoor adjustment formula*—a formula that specifies which variables need to be considered in order to be able to draw a valid conclusion about a cause-and-effect relationship in the case of an *intervention*. We use formal verification to provide computer-verified proofs for all the lemmas and theorems stated in this thesis.

1.1 Formal Verification

Formal verification is a methodology in computer science to prove that a system gives the desired output under all valid inputs. In particular, one use case of formal verification is to write rigorous computer-checked proofs. We explicitly state our assumptions and then progress through the proof with computer-validated steps. Hence, at the end, we have certainty that our proof is correct and that all the assumptions necessary have been explicitly stated. These guarantees of correctness are a notable distinction from classical proof techniques, which are both prone to error and have no guarantee that all assumptions were explicitly acknowledged.

In this thesis, we will be using formal verification for all of our proofs. We create a set of proofs where, without understanding the details of the proof, one can be convinced of their correctness. Additionally, it allows us to start building a foundation for this research that can be extended upon and would result in computer-verified results for science experiments.

We believe that the application of formal verification to these proofs is particularly apt. Scientists are often concerned with how to remove outside influences from their experiments and make their experiments as reliable as possible. That said, many scientists also do not have the math and causal-reasoning background to be able to reason about the factors that may be affecting their experiments or understand all of the proofs presented in this thesis. Hence, we believe that experiment design is an application where the guarantee of correctness provided by formal verification is particularly useful.

We use Rocq [3], a formal-verification system, for all of our proofs. The Rocq proofs written to accompany the lemmas, definitions, and theorems discussed here can be found at https://github.com/giniya5/backdoor_adjustment_formula.

1.2 Background and Related Work

The problem of designing good experiments is well-known and has been studied extensively. Similarly, the concept of an intervention, such as a government policy, is not new—one of the first introductions of this idea was by Haavelmo in 1944 [4,5]. In 1993, interventions were considered in the context of a structural causal model, and the backdoor adjustment formula was proven [6]. It took until 1995 to fully formalize this concept of intervention in structural causal models [7].

In this thesis, we focus on the backdoor adjustment formula and begin the process of formally verifying that it holds. We begin by discussing structural causal models and interventions in order to build the background necessary for the backdoor adjustment formula. Then we discuss the formula itself and present a handful of classical proof approaches that

exist for the formula. We also discuss other work that has focused on formalizing experiment design.

Some efforts to formalize experiment design have focused specifically on structural causal models. There has even been some work done to start formalizing structural causal models in Rocq [8]. We discuss the existing progress in Section 1.2.5 and highlight the way that this thesis is new in using formal verification to prove the backdoor adjustment formula and in trying to provide a formally verified framework and approach to experiment design.

1.2.1 Structural Causal Models

In this thesis, we adopt the structural causal model framework developed by Pearl and others [9]. Structural causal models enable explicit representation of assumptions, graphical reasoning, and the ability to compare observational and interventional expectations.

A structural causal model consists of two main components: a graph and a set of node functions. The graph visually captures the causal relationships between different variables. A node function is assigned to each node and records the effects other nodes have on a specific node, as well as the effects of unobserved outside factors.

In the graph, we represent the various variables, or factors, in our environment as nodes. We add a directed edge between two nodes when there is a causal influence from one variable to the other, to create a directed acyclic graph.

We can see a graph representing a structural causal model in Figure 1.1. Here we can see the various cause-and-effect relationships that could cause one to be late to school. For example, one could wake up late, which might cause them to be late to school. This causal relationship is represented in the graph as an arrow from "waking up late" to "late to school." Alternatively, getting stuck in traffic can also make one late to school, so we have another arrow connecting these two nodes. Similarly, either rain or a traffic accident could cause the traffic jam, so we have two more arrows. We can also see chains of causal effects—for example, a traffic accident could cause a traffic jam, which could then cause one to be late to class.

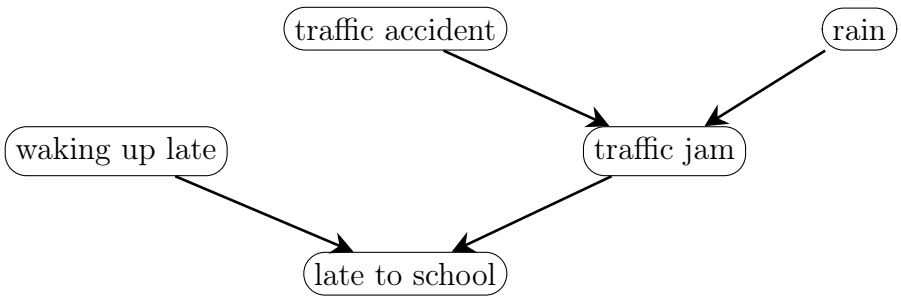


Figure 1.1: A sample structural causal model showing the effects

We also have a node function for each node in the graph. We notice that the edges show causal dependencies, and we can use this information when defining our node functions.

For each node, we say that its value depends on its parents in the graph as well as some unobserved factor (specific to the given node). Hence we can write the random variable

associated with a given node as a function $X_i = f_i(\text{pa}(i), U_i)$, where $\text{pa}(i)$ are the parents of node i and U_i is the unobserved factor.

For example, in Figure 1.1, we can say that the random variable $X_{\text{late to school}}$ depends on the random variables corresponding to its parents, $X_{\text{waking up late}}$ and $X_{\text{traffic jam}}$, as well as an unobserved factor, $U_{\text{late to school}}$. We record this relationship through the following node function,

$$X_{\text{late to school}} = f_{\text{late to school}}(X_{\text{waking up late}}, X_{\text{traffic jam}}, U_{\text{late to school}}).$$

Crucially, we can say that these unobserved factors, U_i, are independent across nodes, since we have captured interdependencies via the edges in this graph. Hence, causal models explicitly capture the relationships between the variables as well as allowing us to note any assumptions we may have made about independence.

Given that in an experiment we usually desire to isolate and test a specific causal relationship between variables (for example, *if* we give a patient this treatment *then* they will recover faster), the ability of this framework to capture causal dependencies makes it a good representation for us to use. The graphical view of a structural causal model makes it easy to work with, while the underlying mathematics also gives us a structured way to record and reason about these dependencies.

1.2.2 Interventions

In the model described in the previous section, we have only worked with observational probabilities. That is, we observe an event and then see how this event affects all the other variables in the model. While we often want to study such observational effects, sometimes we are interested in the case where we instead intervene on a variable rather than just observing it. For example, a policy is a form of intervention. Often, once a policy is implemented, it is a blanket rule for the population. That is, rather than making a decision that is influenced by some factors, all of these influences are removed, and the decision is made for the general population.

Note that intervening on variables, instead of just observing them, can result in substantially different outcomes. For example, consider the probability of there being a fire if smoke has been observed, $\Pr(\text{fire} \mid \text{smoke})$. We would expect that this probability is high—when one sees smoke, one also expects a fire. On the other hand, consider the probability of a fire if one has intervened to create the smoke. We denote this probability as $\Pr(\text{fire} \mid \text{do}(\text{smoke}))$, where the do denotes the fact that we have created the smoke, rather than observing it. We would expect this probability to be fairly low. On average, the probability of a fire $\Pr(\text{fire})$ is fairly low. If one has caused the smoke (without making a fire), then there is no reason to expect that there is a fire around, so $\Pr(\text{fire} \mid \text{do}(\text{smoke}))$ will stay fairly low. In fact, we notice that it seems that $\Pr(\text{fire} \mid \text{do}(\text{smoke}))$ is more likely to be equal to $\Pr(\text{fire})$ than to be equal to $\Pr(\text{fire} \mid \text{smoke})$.

It is not always the case that intervening on a variable causes different outcomes to observing it. Consider instead the effect of doing practice tests on a test score. We expect (or at least hope!) that $\Pr(\text{good score on test} \mid \text{did practice tests})$ is high. Now consider a situation where one is forced to do the practice tests. The probability $\Pr(\text{good score on test} \mid \text{do}(\text{did practice tests}))$ is probably still fairly high—it does not really matter why someone does the

practice tests, rather it is more important that learnt from doing them. It seems more likely that Pr(good score on test | do(did practice tests)) is similar to Pr(good score on test | did practice tests), rather than Pr(good score on test), unlike the fire and smoke example.

The observation that intervening on a variable can, but does not always, lead to substantially different outcomes than just observing a variable makes this distinction regarding observational and interventional probabilities an important one. To be able to reason about this distinction, Pearl formally defines the concept of an *intervention* in a structural causal model as follows [9]:

Definition 1.2.1. *Given a structural causal model, we can perform an atomic intervention setting the value of a variable V_i to be the constant v. In the graph corresponding to this model, we must also remove any incoming edges to node i, corresponding to V_i, since no other variables influence the values of $V_i = v$ any longer. We denote this intervention as* $\text{do}(V_i = v)$.

We note that policies are almost always interventions, since they force people to behave in specific ways. However, sometimes one must decide whether to implement a policy with only access to observational data and probabilities. By understanding the underlying structural causal model, it is possible to understand the relationships between interventional and observational probabilities. In this thesis, we focus on a single relationship between interventional and observational probabilities, the backdoor adjustment formula.

1.2.3 Backdoor Adjustment Formula

In order to be able to explain the backdoor adjustment formula and when it holds, we must first continue our exploration of structural causal models. The dependencies between variables are what determines how we can construct experiments to learn things about the interventional probability and its relation to observational probabilities.

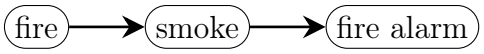
In causal models, there are three basic node structures that we are particularly concerned with:

Definition 1.2.2. *Let $G = (V, E)$ be a structural causal model, where $a, b, c \in V$.*

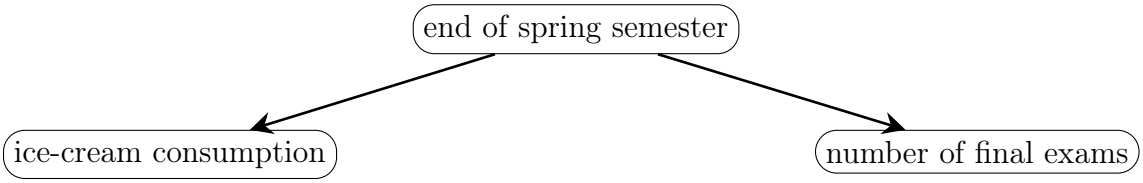
- *b is a* mediator *of a and c if $(a, b) \in E$ and $(b, c) \in E$ or if $(b, a) \in E$ and $(c, b) \in E$.*
- *b is a* confounder *of a and c if $(b, a) \in E$ and $(b, c) \in E$*
- *b is a* collider *of a and c if $(a, b) \in E$ and $(c, b) \in E$.*

We depict these three node structures in Figure 1.2. In Figure 1.2(a), we see that smoke is a mediator of fire and a fire alarm. The fire causes smoke, and then the smoke triggers the fire alarm. Note that a fire does not directly trigger a fire alarm, which is represented in the graph by the absence of an edge between fire and fire alarm.

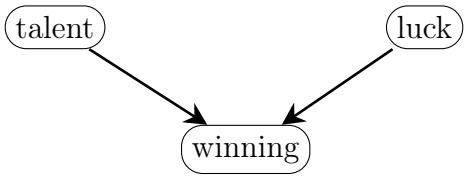
Figure 1.2(b) shows a confounder, the end of the spring semester. As we approach the end of the semester and it gets warmer, people are likely to consume more ice cream. Similarly, many students have finals at the end of the semester. Thus, one may observe that when ice cream consumption increases, the number of final exams increases. This correlation is not



(a) A structural causal model with a mediator (smoke)



(b) A structural causal model with a confounder (end of spring semester)



(c) A structural causal model with a confounder (winning)

Figure 1.2: A mediator, confounder and collider

due to any relationship between the ice cream and exams but is due to the confounder, which is that it is the end of the spring semester. Thus there is no edge between ice cream and exams.

Figure 1.2(c) shows an example of a collider, winning some competition. Both talent and luck can help one win, but there is no relationship between talent and luck themselves, so there is no edge between them. However, we can create the appearance of such a relationship by conditioning on people who have won. It may appear that if someone is talented, they are not likely to be lucky and vice versa. This relationship may be because people probably need either talent or luck to win, so we are only looking at the set of people who are either talented or lucky.

In these three examples, there were variables that were not connected by edges. Some of these variables were clearly related, like fire and fire alarms, while some did not seem very related, like luck and talent. When identifying causal relationships, a concept that is interesting to us is which variables affect each other and which ones do not, regardless of whether they have an edge connection. In other words, we desire some concept of independence that can be read from the graph. To capture this notion of independence, Pearl introduces a graphical criterion called *d-separation* [9], which we define in Definition 1.2.3. It has been formally verified that semantically d-separation is equivalent to independence between nodes [8].

Definition 1.2.3. *A path p is said to be* d-separated *(or* blocked*) by a set of nodes Z if and only if*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or*

2. *p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.*

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

Now that we have this graphical notion of d-separation, we are able to use it to define the *backdoor criterion*. The backdoor criterion will give us a set of conditions that result in us being able to learn about interventional probabilities from only observational probabilities, as seen in Definition 1.2.4.

Definition 1.2.4 (Backdoor Criterion)**.** *A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:*

1. *no node in Z is a descendant of X_i; and*

2. *Z blocks every path between X_i and X_j that contains an arrow into X_i.*

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the backdoor criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

We call a path between X_i and X_j that contains an arrow into X_i a *backdoor path*, since these are paths that connect X_i and X_j and have the potential to influence X_i. This idea of backdoor paths is what gives the criterion its name.

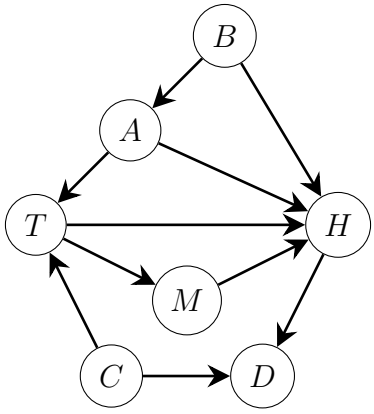


Figure 1.3: An example structural causal model

Consider the structural causal model shown in Figure 1.3. We notice that in this case there are three backdoor paths between T and H:

1. $T - A - B - H$

2. $T - A - H$

3. $T - C - D - H$

Notably, the paths $T - H$ and $T - M - H$ are not backdoor paths since they do not involve edges incoming into T. Now, we consider how to block these backdoor paths. Observe that the third backdoor path is already blocked since it contains a collider (D). We can block both of the first two paths by conditioning on the variable A. A is a mediator in the first path and a confounder on the second path, so in both of these cases A blocks the path. Hence we have a set $Z = \{A\}$ which satisfies the backdoor criterion relative to (T, H).

The set $Z = \{A, B, C, D\}$ also satisfies the backdoor criterion relative to (T, H). In this case, A still blocks both of the first two paths. The third path is no longer blocked by D since D is a collider, and we are now conditioning on it. However, it is blocked by C now, since C is a confounder on the path.

On the other hand $Z = \{B, C\}$ does not satisfy the backdoor criterion relative to (T, H). In this case, the second backdoor path, $T - A - H$, is not blocked.

We can now use the concept of the backdoor criterion to state the backdoor adjustment formula in Theorem 1.2.5.

Theorem 1.2.5 (Backdoor Adjustment Formula)**.** *If a set of variables Z satisfies the backdoor criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$\Pr(y|do(x)) = \sum_z \Pr(y|x, z) \Pr(z).$$

We once again consider the structural causal model shown in Figure 1.3. We noted that $Z = \{A\}$ satisfied the backdoor criterion relative to (T, H). We can then apply the backdoor adjustment formula, to find that in the structural causal model shown in Figure 1.3 the interventional probability can be computed from the observational probability, and the relationship between them is

$$\Pr(H = h|\text{do}(T = t)) = \sum_a \Pr(H = h|T = t, A = a) \Pr(A = a).$$

In this thesis, all of our work revolves around the backdoor adjustment formula. We first prove that it holds for certain small examples in Chapter 2. We then move on to prove it in the general case, but with a couple of extra assumptions, in Chapter 3. We also show the applicability of the backdoor adjustment formula in science by examining papers where experiments were conducted in Chapter 4.

1.2.4 Proofs of the Backdoor Adjustment Formula

There are a number of different proofs of the backdoor criterion, although in general, the level of detail in these classical proofs is not quite enough to be able to transform them into a formally verified proof without some extra work. The backdoor criterion was first proven by Pearl in [6]. While we considered formalizing the proof approach used there, the proof is somewhat indirect—it works by introducing another node that captures the concept of an intervention. The proof them uses some independence relations to reason about the new graph. Formalizing such a proof approach creates an extra step—we would need to show that the new formulation with an extra node is equivalent to capturing the graph with and

without a do-intervention. We would also need to formalize this new node in Rocq, capturing its ability to exist in two states and the effects it has on all the other variables. The creation of an extra node, as well as the difficulty in proving some of the independence statements required, made the most common classical proof approach seem like a suboptimal choice for our formally verified proof.

Another way to prove the backdoor adjustment formula is by using the second rule of do-calculus, which also relates interventional and observational probabilities. However, this proof approach does not actually simplify our proof since the second rule of do-calculus is slightly more general than the backdoor criterion.

The proof of the backdoor adjustment formula that we ended up using as the basis for our formally verified proof is presented in *Causality* [9] as an alternate proof for the backdoor adjustment formula. This proof proves the backdoor adjustment formula directly. Additionally, unlike in other approaches, this proof has fewer subtleties, which must all be fleshed out when formally verifying a proof.

1.2.5 Existing Formalizations of Experiment Design

There is literature on experiment design, but it usually has a focus on how to conduct an experiment in practice rather than focusing on the underlying mathematical principles and assumptions. Many resources focus on providing concrete advice for experiment designers [10], creating frameworks for more reliable experiments [11], or studying past experiments and noting the inaccuracy of results [12]. There is even some work that allows for interactive reasoning about causal graphs, such as Causal Fusion [13]. However, relatively little work provides a mathematically rigorous treatment of experiment designs and explicitly identifies the assumptions required to infer causal effects from observed data. This thesis seeks to fill that gap by using formal verification to provide a rigorous computer-verified framework grounded in causal inference.

This thesis is not the first place where we use formal verification to provide guarantees about structural causal models. Work done by Zhang [8] began this process, by defining a structural causal model in Rocq and then formally verifying some properties and theorems about structural causal models. While we do not explicitly build on this work, our work is closely related. We believe that eventually the combination of both of these works would allow for a formally verified way to work with causal models and design experiments that follow the criterion required for the backdoor adjustment formula, allowing scientists to reach conclusions that are mathematically sound and to state explicitly all assumptions that were made.

Chapter 2

Backdoor Adjustment Formula for Example Graphs

This chapter verifies the backdoor criterion for specific small causal models. To begin, we discuss our representations of node functions and the do-operator in Rocq. Then, we verify the backdoor adjustment formula on the simplest possible graph, which consists of only two nodes. Finally, we verify the backdoor adjustment formula on slightly more interesting but still simple graphs consisting of three nodes and containing either a mediator, a confounder, or a collider.

2.1 Node Functions and the Do-Operator in Rocq

When we consider the random variables associated with a node in the structural causal model, we restrict them to be discrete random variables that draw from finite sample spaces and have finite domains. This choice is made in order to make it easier to work with the node functions in Rocq. Specifically, the infotheo library [14], which defines many critical concepts including random variables and probabilities, primarily works with discrete random variables. Many of the lemmas proven in the library are only applicable in the discrete case, and thus, we meet many more roadblocks when working with continuous random variables. Additionally, we are able to use summation rather than integration in the discrete case, which further simplifies our Rocq code.

While an experiment may have random variables that are not discrete, in many cases it does make sense to assume that the variables are discrete. For example, consider studying whether taking a medication means that a patient recover from a disease. We notice that often taking medication and recovery are both binary variables—either the patient recovers, or they do not. Studies also often assume that their random variables are discrete and finite in order to simplify their models of the world. In fact, this assumption is one that is rather common in causal inference.

We also need a representation of the do-operator in Rocq. The standard notation used is $\Pr(y|\text{do}(x))$, which is shorthand for $\Pr(Y = y|\text{do}(X = x))$, where X and Y are random variables, while x and y are specific values that the random variables take, and we are setting $X = x$ via intervention. This notation is somewhat difficult to work with since the $\text{do}(X = x)$

term is not a random variable or a standard probabilistic concept, unlike X and Y. Instead $\text{do}(X = x)$ represents a change we have made to our structural causal model and all of the random variables in it.

When we write $\text{do}(X = x)$ we are modifying Y (or any random variable in a probability expression containing $\text{do}(X = x)$), by forcibly setting the value of the random variable X to x. Thus, despite using the same notation Y in both $\Pr(Y = y|\text{do}(X = x))$ and $\Pr(Y = y|X = x)$, they are not the same random variable whenever Y depends on X. We exploit the fact that these are two different random variables to rewrite probabilities containing an intervention $\text{do}(X = x)$ as probabilities on new post-intervention random variables.

In a structural causal model, a random variable is a function of some other random variables. In the case of an intervention $\text{do}(X = x)$, we can update all of our node functions to a post-intervention state. Then we can write all of our dependencies as before but replace any dependency on X with x and have node functions depend on the post-intervention version of other node functions. We use the notation $Y_{X=x}$ to represent the node function after the intervention $\text{do}(X = x)$.

Remark 2.1.1. For a structural causal model, containing at least two nodes X and Y, we can write

$$\Pr(Y = y|\text{do}(X = x)) = \Pr(Y_{X=x} = y)$$

where $Y_{X=x}$ is a random variable corresponding to Y in the model post-intervention.

In the Rocq implementation, we work with the concept of these modified random variables $Y_{X=x}$, rather than working with the concept of the do-operator as something to condition on. By using the modified-random-variable version of a do-intervention, we are able to define all of our random variables and their post-intervention counterparts explicitly. The infotheo library treats all random variables as functions, making such a definition fairly straightforward, since we can simply change the function input from a random variable X to the intervened value x.

2.2 Two-Node Model

The simplest model we can consider is one that consists of only two variables/nodes. Let us call these nodes T and H, corresponding to treatment and health. We consider the effect of treatment on health. We note that since there are no other nodes in the graph, we are also assuming that there are no other variables that influence the treatment and health outcomes and further, that treatment and health do not influence other variables. The graph corresponding to this structural causal model can be seen in Figure 2.1.

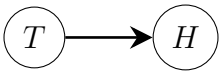


Figure 2.1: A two-node structural causal model

We now consider the backdoor criterion for this graph. Recall that to satisfy the backdoor criterion, we need a set Z that satisfies both of the conditions in Definition 1.2.4—specifically, that Z blocks all backdoor paths between T and H and that Z contains no descendants of T.

We claim that the set $Z = \varnothing$ satisfies the backdoor criterion. Since there are no paths containing chains or forks, we do not have any backdoor paths in this graph. As it is the empty set, it also necessarily contains no descendants of T. Thus, we can apply the backdoor adjustment formula for (T, H) with the set $Z = \varnothing$, to find that

$$\Pr(H|\text{do}(T = t)) = \sum_z \Pr(H|T = t, Z = z) \Pr(Z = z) = \Pr(H|T = t).$$

This equation becomes the central lemma that we verify using Rocq. In order to prove this equation holds, we need some further assumptions. One of these is an underlying assumption of any structural causal model, which is that the unobserved factors are mutually independent. The second condition is marginally more interesting and is a condition that was only added once we tried to prove this lemma in Rocq and found that it was necessary. We require that we intervene by setting T to a value t that can be achieved without intervention. In other words, we require $\Pr(T = t) \neq 0$.

While the non-zero condition is not explicitly stated in the backdoor adjustment formula, we notice that this positivity condition is actually a very natural requirement, and one that is implicitly present in classical proofs. The probability $\Pr(H = h|T = t)$ is not well-defined when $\Pr(T = t) = 0$, so it does not make sense to reason about the zero case. Another way to think about this condition is that if $\Pr(T = t) = 0$, this means we cannot observe this event unless we intervene to set $T = t$. But, if this is the case, it is natural that we cannot learn anything about such an intervention from observational data since there is no observational data about this event.

We state our lemma, with these two conditions, in Lemma 2.2.1.

Lemma 2.2.1. *Consider the two-node structural causal model $T \to H$. If all unobserved factors are mutually independent and $\Pr(T = t) \neq 0$, then*

$$\forall h.\ \Pr(H = h|do(T = t)) = \Pr(H = h|T = t).$$

To prove Lemma 2.2.1, we need to set up a probability distribution over a sample space. We have two nodes T and H. The nodes have unobserved factors, which can take values from the sets UT and UH. We set up our sample space to be $UT \times UH$, since then we can consider our graph under any set of unobserved factors. We can then define a probability distribution P over our sample space. We also define a space of outcomes from our random variables *outcomes*, to later be able to define random variables that map from the probability distribution P to a value in *outcomes*. We take *outcomes* to be the union of all possible outcomes of all random variables. In this way, we do not lose any generality, despite only defining one outcomes space for both of our random variables.

We discuss the way that we encode these variables in Rocq. For this two-node example we show a lot of our definitions and lemmas. For later examples we often omit Rocq code where it looks very similar to previously shown Rocq code.

```
Context {R : realType}.
Variables (UT UH : finType).
Variable (outcomes : finType).
Variable P : R.-fdist (UT * UH).
```

We can use the structural causal model to determine what the node functions should be. First, we define some functions f_T and f_H. These correspond to the functions at all nodes and take some values from the parents and unobserved variables as inputs, before outputting values for nodes. While we define the inputs and outputs of these functions, we crucially never define these functions themselves, since we do not know what the functions are from a structural causal model, only the dependencies. Thus, this representation allows us to represent any structural causal model that has a graphical representation as in Figure 2.1.

In our specific two-node example, we notice that from the graph, we know that T depends only on its unobserved term, while H depends on T and its unobserved term. Thus, we shall have a function f_T for the vertex T that will take an unobserved term (of type UT) as input and output something in *outcomes*. In the meantime, f_H will take both an unobserved term (of type UH) and the outcome from T (which will be of type *outcomes*). We use the inputs and outputs to define `fT`=f_T and `fH`=f_H in Rocq.

```
Variable fT : UT -> outcomes.
Variable fH : UH -> outcomes -> outcomes.
```

We also define random variables corresponding to the unobserved terms, in order to be able to write our assumption about unobserved terms being independent. Since we have defined the unobserved terms to be the sample space, the definitions for the unobserved random variables are simple:

```
Let UTRV : {RV P -> UT} := fun u => u.1.
Let UHRV : {RV P -> UH} := fun u => u.2.
```

We then move on to creating random variables that correspond to the nodes in our graph. Notice that from our graph we can determine our node functions. Namely, we have $T = f_T(U_T)$ and $H = f_H(U_H, T)$. In the case of an intervention $do(T = t)$, we have $H_{T=t} = f_H(U_H, t)$. We use the notation T=`Tnodefn`, H=`Hnodefn` and $H_{T=t}$=`Hnodefnint` to define these node functions in Rocq.

```
Let Tnodefn : {RV P -> outcomes} :=
  fun (u : UT * UH) => fT (UTRV u).
Let Hnodefn : {RV P -> outcomes} :=
  fun (u : UT * UH) => fH (UHRV u) (Tnodefn u).
Let Hnodefnint (t : outcomes) : {RV P -> outcomes} :=
  fun (u : UT * UH) => fH (UHRV u) t.
```

Finally, with all of this set up, we can state Lemma 2.2.1 in Roq.

```
Lemma two_var_backdoor_adjustment: forall (t : outcomes),
  P |= UHRV _|_ UTRV ->
  ‘Pr[ Tnodefn = t] != 0 ->
  forall a, ‘Pr[ (Hnodefnint t) = a] = ‘Pr[ Hnodefn = a | Tnodefn = t].
```

The lemma states that if the random variables corresponding to the unobserved factors of T and H are independent (`P| = UHRV _|_ UTRV`), and we intervene by setting T to some value t that is reachable by T without intervention (`‘Pr[ Tnodefn = t ] != 0`), then for all a, the interventional and observational probabilities of H given T are equal.

We prove Lemma 2.2.1 by proving two separate lemmas: one that lets us take the independence of unobserved variables and then prove something about the independence of the node functions, and a second that lets us start with the independence of the node functions and then lets us prove that the probability distributions under the intervention $T = t$ or the observation that $T = t$ are equivalent.

The first of these lemmas is as follows:

```
Lemma indep_implication: forall t,
  P |= UHRV _|_ UTRV ->
  P |= (Hnodefnint t) _|_ Tnodefn.
```

In other words, this lemma says that if we know that the unobserved factors are mutually independent (which is always true for a causal model), then we can concluded that $H_{T=t} \perp T$.

This lemma is true because $H_{T=t}$ depends only on one random variable, which is $UHRV$. We have removed the dependence it had on T through our do-intervention. Similarly, T only depends on the random variable $UTRV$. If one transforms two random variables, any independence conditions are maintained, and thus we are able to show that this lemma holds.

The second of these lemmas says the following:

```
Lemma prob_version: forall t,
  P |= (Hnodefnint t) _|_ Tnodefn ->
  ‘Pr[ Tnodefn = t] != 0 ->
  forall a, ‘Pr[ Hnodefn = a | Tnodefn = t] = ‘Pr[ (Hnodefnint t) = a].
```

The independence between $H_{T=t}$ and T allows us to rewrite $\Pr(H_{T=t} = a)$ as $\Pr(H_{T=t} = a|T = t)$, since independence guarantees that these two probabilities are equal. Then, we consider the comparison between $\Pr((H_{T=t}) = h|T = t)$ and $\Pr(H = h|T = t)$. By unfolding the definition of conditional probabilities, we are able to consider whether $\Pr((H_{T=t}) = h, T = t)$ and $\Pr(H = h, T = t)$ are equal, which is true due to our definitions of the node functions.

2.3 Mediator

Having considered a graph with two nodes, we now consider graphs with three nodes. We consider the three structures we introduced in Definition 1.2.2. We start with a three-node graph containing a mediator, as shown in Figure 2.2.

We notice that if we consider the backdoor criterion, in this case, it is once again satisfied if we take the set $Z = \varnothing$ relative to (T, H). Thus, the lemma we are proving looks very similar to Lemma 2.2.1. We state our new lemma as Lemma 2.3.1.

Lemma 2.3.1. *Consider the three-node structural causal model shown in Figure 2.2. If all unobserved factors are mutually independent and we chose an intervention value t such that $\Pr(T = t) \neq 0$, then*

$$\forall h.\ \Pr(H = h|do(T = t)) = \Pr(H = h|T = t).$$

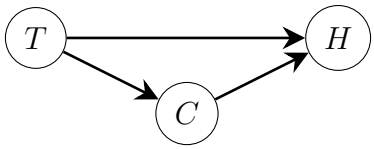


Figure 2.2: A three-node structural causal model with a mediator

As before, we define the sample space for this graph to be the unobserved variables, so $UC \times UT \times UH$, and define a probability distribution P over this space. We define random variables for our unobserved terms, which we call $UTRV$, $UCRV$ and $UHRV$. We can then define our node functions, just as in Section 2.2, but with extra dependencies due to the new node C. As before, we first must define some functions f_T, f_C and f_H, which are unknown to us but take specific inputs. Then, we can use these functions to define node functions that match Figure 2.2: $T = f_T(U_T)$, $C = f_C(U_C, T)$, and $H = f_H(U_H, C, T)$.

```
Let Tnodefn : {RV P -> outcomes} :=
  fun (u : UT * UC * UH) => fT (UTRV u).
Let Cnodefn : {RV P -> outcomes} :=
  fun (u : UT * UC * UH) => fC (UCRV u) (Tnodefn u).
Let Hnodefn : {RV P -> outcomes} :=
  fun (u : UT * UC * UH) => fH (UHRV u) (Cnodefn u) (Tnodefn u).
```

We also define these node functions under intervention, by replacing every instance of the variable `Tnodefn` with the intervention value, t, and working with the intervention version of the functions for all of the other node functions as well. Notably, in this case the node function for H depends on C, and this dependence exists even after intervention, however, $H_{T=t}$ should depend on the post-intervention version of C, $C_{T=t}$. Thus, we get the node function $H_{T=t} = f_H(U_H, C_{T=t}, t)$. Similarly, in our Rocq code, `Hnodefnint` now calls `Cnodefnint` instead of `Cnodefn`.

```
Let Cnodefnint (t : outcomes) : {RV P -> outcomes}:=
  fun (u : UT * UC * UH) => fC (UCRV u) t.
Let Hnodefnint (t : outcomes) : {RV P -> outcomes}:=
  fun (u : UT * UC * UH) => fH (UHRV u) (Cnodefnint t u) t.
```

We now write Lemma 2.3.1 in Rocq, which is the core lemma we focus on proving in this section. We use similar notation as in Section 2.2, with the exception that now we need mutual independence between all three unobserved random variables, which is conveyed by `mutual_indep_three UHRV UTRV UCRV`.

```
Lemma three_var_mediator_backdoor_adjustment: forall t,
  mutual_indep_three UHRV UTRV UCRV ->
  ‘Pr[ Tnodefn = t] != 0 ->
  forall a, ‘Pr[ Hnodefn = a | Tnodefn = t] = ‘Pr[ (Hnodefnint t) = a].
```

The proof of Lemma 2.3.1 is very similar to the two-variable case. We do some extra work since when we consider the node function of H, we also end up having to consider the node C. That said, we are once again able to prove it by showing first that the mutual independence of $UTRV$, $UCRV$ and $UHRV$ means that $(H_{T=t}) \perp T$. Since T does not depend on C, this proof barely changes. Then, from there, we can conclude that the probability distributions are equivalent, regardless of whether we intervene or observe the value of T. In this part, our definitions for the node functions differ from Section 2.2; however, the proof is structurally the same. In other words, we show that, given any value of t, for all values of h, $\Pr(H = h|T = t) = \Pr((H_{T=t}) = h)$, and hence we prove Lemma 2.3.1.

2.4 Confounder

We move onto the three-variable case where we have a confounder. This structural causal model is depicted in Figure 2.3.

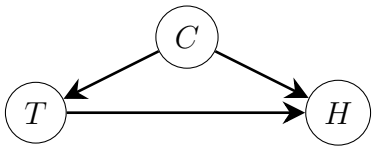


Figure 2.3: A three-node structural causal model with a confounder

Unlike the previous examples, the set $Z = \varnothing$ relative to (T, H) does not satisfy the backdoor criterion, since we leave the backdoor path $T \leftarrow C \rightarrow H$ unblocked. In order to block this path, we must actually include C in our conditioning set Z. In fact, $Z = \{C\}$ does satisfy the backdoor criterion, as C is not a descendant of T, and we have now blocked all backdoor paths. Since we now have a nonempty set Z, we have a slightly different version of the lemma from the last two examples, as seen in Lemma 2.4.1.

Lemma 2.4.1. *Consider the three-node structural causal model shown in Figure 2.3. If all unobserved factors are mutually independent and we chose an intervention value t such that for all $c \in range(C), \Pr(T = t|C = c) \neq 0$, then we get that*

$$\forall h.\ \Pr(H = h|do(T = t)) = \sum_{c \in range(C)} \Pr(H = h|T = t, C = c) \Pr(C = c).$$

We note that we have a summation over $c \in range(C)$ in Lemma 2.4.1. We sum over the range of C rather than the domain of C, since in our probability term we consider whether the random variable C is equal to c, rather than considering the output of C when the input is c.

Observe that the nonzero condition in Lemma 2.4.1 has changed slightly from our previous nonzero conditions—specifically, instead of just requiring that we are intervening on a value of t that is possible to achieve without intervention, we now require that for all possible outcomes of C, it is possible to achieve t without intervention. Within Rocq, this statement is equivalent to saying that $\Pr(T = t, C = c)$ is not zero, which would then match the condition

in Lemma 2.2.1 and 2.3.1, since we require that all of our probabilities are well-defined, and $\Pr(H = h|T = t, C = c)$ is not well-defined when $\Pr(T = t, C = c) = 0$.

We are able to approach this problem very similarly to before. However, unlike before, we need to handle the summation over $c \in range(C)$. We note that if

$$\Pr(H = h|T = t, C = c) = \Pr(H_{T=t} = h|C = c), \tag{2.1}$$

then we can use the law of total probability, given in Theorem 2.4.2, to complete the proof.

Theorem 2.4.2 (Law of Total Probability)**.** *Given two random variables A and B. If a is a value the random variable A can take, then*

$$\Pr(A = a) = \sum_{b \in range(B)} \Pr(A = a|B = b) \Pr(B = b).$$

The law of total probability is already formally verified in Rocq in the infotheo library. We are able to use it show that

$$\sum_{c} \Pr(H_{T=t} = h|C = c) \Pr(C = c) = \Pr(H_{T=t} = h)$$

as long as equation 2.1 is true. Thus, we turn our attention to proving that (2.1) is true. We notice that this equation resembles the equations in Lemmas 2.2.1 and 2.3.1, except with an added condition on C. It turns out that we can use many of the same ideas to prove (2.1).

We define our node functions in Rocq in a similar fashion to that presented in the previous two sections. To start, we prove that the independence of our node functions can be derived from the independence of the unobserved terms.

```
Lemma mut_unobs_indp_cond_indp_wo_inj: forall t,
  mutual_indep_three UHRV UTRV UCRV ->
  (Hnodefnint t) _|_ Tnodefn | Cnodefn.
```

Unlike before, this independence statement does not follow as easily from the mutual independence of unobserved term, since there are shared dependencies between $H_{T=t}$ and T. However, this shared dependence is entirely on C, and we are conditioning on C. Thus, we are able to show that we can conclude that $H_{T=t} \perp T|C$ from the mutual independence of the unobserved terms.

We then use a lemma similar to the one we used in the case of the two-node graph, just with the extra conditioning on C, as follows:

```
Lemma doint_equiv_with_confounder_prob: forall t c,
  (Hnodefnint t) _|_ Tnodefn | Cnodefn ->
  `Pr[ Tnodefn = t | Cnodefn = c] != 0 ->
  (forall h, `Pr[ Hnodefn = h | [% Tnodefn, Cnodefn] = (t, c)]
      = `Pr[ (Hnodefnint t) = h | Cnodefn = c]).
```

Just as in the two-node graph example, we can prove Lemma 2.4.1 by rewriting $\Pr(H_{T=t}|C = c)$ as $\Pr(H_{T=t}|C = c, T = t)$ using our independence condition. We then consider the definitions of our node functions to find that, after conditioning on both C and T, the node

function for H returns the same value regardless of whether there was an intervention on T or not. In other words, we get that $\Pr(H_{T=t} = h|C = c, T = t)$ and $\Pr(H = h|C = c, T = t)$, which completes our proof.

Putting these two lemmas together proves equation 2.1, leading us to the following lemma:

```
Lemma three_var_confounder_backdooor_adjustment_eq: forall t c,
  mutual_indep_three UHRV UTRV UCRV ->
  ‘Pr[ Tnodefn = t | Cnodefn = c] != 0 ->
  (forall h, ‘Pr[ (Hnodefnint t) = h | Cnodefn = c] =
      ‘Pr[ Hnodefn = h | [% Tnodefn, Cnodefn] = (t, c)]).
```

We can then use the lemma above and the law of total probability to prove the backdoor criterion in the case of a confounder. We prove the following Rocq lemma, which is equivalent to Lemma 2.4.1.

```
Lemma three_var_confounder_backdoor_adjustment: forall t,
  mutual_indep_three UHRV UTRV UCRV ->
  (forall c,  ‘Pr[ Tnodefn = t | Cnodefn = c] != 0) ->
  (forall h, ‘Pr[ (Hnodefnint t) = h] =
      \sum_(c in outcomes) (‘Pr[Hnodefn = h | [% Tnodefn, Cnodefn] = (t, c)]
          * ‘Pr[ Cnodefn = c])).
```

2.5 Collider

The final example of a structure introduced in Definition 1.2.2 is a collider. We can see a three-node structural casual model with a collider C in Figure 2.4.

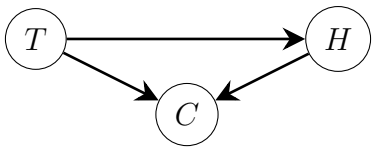


Figure 2.4: A three-node structural causal model with a collider

In this case we once again notice that the backdoor criterion is satisfied by $Z = \varnothing$ relative to (T, H). Thus, we find we have a similar lemma to the one we had in the two-node and mediator cases, namely, Lemma 2.5.1, as follows:

Lemma 2.5.1. *Consider the three-node structural causal model shown in Figure 2.4. If all unobserved factors are mutually independent and we chose an intervention value t such that $\Pr(T = t) \neq 0$, then*

$$\forall h.\ \Pr(H = h|do(T = t)) = \Pr(H = h|T = t).$$

We define our node functions in Rocq as before. We prove Lemma 2.5.1 by first using mutual independence of unobserved factors to find that the node functions are independent.

We then use the independence of the node functions to find that the probability distributions are equivalent. We note that since neither T nor H depend on C, this proof is actually virtually unchanged from the proof presented in Section 2.2. The lemma that is proven in Rocq is copied below.

```
Lemma three_var_collider_backdoor_adjustment: forall t,
  mutual_indep_three UHRV UTRV UCRV ->
  ‘Pr[ Tnodefn = t] != 0 ->
  forall a, ‘Pr[ (Hnodefnint t) = a] = ‘Pr[ Hnodefn = a | Tnodefn = t].
```

2.6 Implementation Details

It is important to note the Rocq axioms that we used in order to prove the lemmas in this chapter. Rocq is built on the Calculus of Inductive Constructions [15], so we have used the inference rules associated with it. Beyond that, we also use the axioms

- `boolp.propositional_extensionality`,
- `boolp.functional_extensionality_dep` and
- `boolp.constructive_indefinite_description`.

These axioms are consistent with Rocq and are axioms that are taken for granted in classical proofs. For example, we have

```
boolp.propositional_extensionality :
  forall P Q : Prop, P <-> Q -> P = Q.
```

In other words, if there are two propositions that imply each other, then they are the same proposition. In the case of a confounder, presented in Section 2.4, we also used the axiom `classic`. Beyond these axioms, all of our proofs were fully proven, and we did not have to make any more assumptions beyond those explicitly stated in lemmas.

Chapter 3

Backdoor Adjustment Formula in the General Case

We move on to proving the backdoor adjustment formula in the general case. It seems natural to try to mirror the approach we used when proving the small examples in the previous chapter. However, we find that the final step, which involved showing that $Pr(H_{T=t} = h|T = t, Z = z) = P(H = h|T = t, Z = z)$ (or that $\Pr(H_{T=t} = h|T = t) = \Pr(H = h|T = t)$ in the case where $Z = \varnothing$) is difficult. In Chapter 2, we considered the definition of the node functions H and $H_{T=t}$ to prove the equalities above. Since we do not know what the node functions are in the general case, we cannot use this approach. It turns out that it is somewhat difficult to get a relation between the probability with and without intervention.

The way we approach relating the interventional and observational probabilities is by considering the joint probability distribution of the two structural causal models. We simplify these distributions with graph factorization and then rearrange them to get a probabilistic equation which contains the probabilities of our outcome H both before and after a do-intervention. This equation allows us to make the final jump between the node functions $H_{T=t}$ and H.

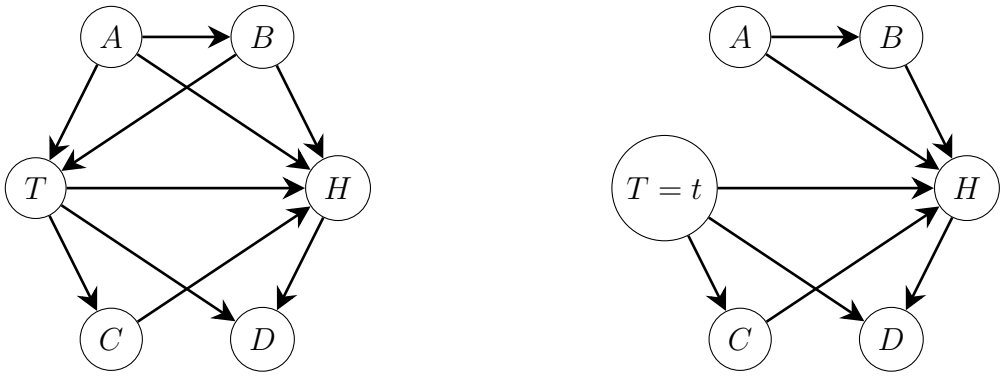
Even using the joint probability distributions, it is still somewhat difficult to prove that $\Pr(H_{T=t} = h|T = t) = \Pr(H = h|T = t)$. Instead, we use a different approach in this chapter, first proving the *parental adjustment formula* and then using that to prove the backdoor criterion.

3.1 Joint Probability Distributions and Graph Factorization

The joint probability distribution for a graph measures the simultaneous probability of all random variables (which correspond to the nodes) in the graph taking specific values. Consider a graph $G = (V, E)$ with n vertices V_i, $i \in [1, n]$. Each node has a node function, which is a random variable X_i, $i \in [1, n]$. Then, we can denote this joint probability distribution as

$$\Pr(X_1, X_2, \ldots, X_n).$$

In the case of a directed acyclic graph, we can factorize this joint distribution as follows:



(a) A structural causal model with no intervention

(b) A structural causal model with intervention $T = t$

Figure 3.1: Two structural causal models, one with and one without intervention

Lemma 3.1.1 (Graph Factorization for Directed Acyclic Graphs)**.** *Given a directed acyclic graph $G = (V, E)$ with n vertices. Let each vertex correspond to a random variable X_i. Then, we can factorize the joint probability distribution as follows,*

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} Pr\left(X_i = x_i | \mathrm{PA}(X_i) = pa(X_i)\right),$$

where $\mathrm{PA}(X_i)$ is the set of random variables corresponding to the parents of the node V_i, and $pa(X_i)$ are the values the elements in this set take.

By construction, a structural causal model is always a directed acyclic graph. Thus, we can use the above factorization to simplify out joint probability distribution.

Consider the example structural causal model shown in Figure 3.1(a). We use Lemma 3.1.1 to factorize this graph. Let the random variable associated with node I be X_I. Then, we find that the joint probability distribution for this graph is

$$\begin{aligned}
&\Pr(X_T = x_t, X_H = x_h, X_A = x_a, X_B = x_b, X_C = x_c, X_D = x_d) \\
&= \prod_{i \in \{T,H,A,B,C,D\}} \Pr(X_i = x_i | \mathrm{PA}(X_i) = \mathrm{pa}(X_i)) \\
&= \Pr(X_T = x_t | X_A = x_a, X_B = x_b) \cdot \Pr(X_H = x_h | X_A = x_a, X_B = x_b, X_C = x_c, X_T = x_t) \\
&\quad \cdot \Pr(X_A = x_a) \cdot \Pr(X_B = x_b | X_A = x_a) \cdot \Pr(X_C = x_c | X_T = x_t) \\
&\quad \cdot \Pr(X_D = x_d | X_H = x_h, X_T = x_t).
\end{aligned}$$

We can also consider the same graph once we intervene on T, to set $T = t$. The structural causal model after intervention is shown in Figure 3.1. All of the incoming edges to T have been removed, since they no longer influence the value of X_T.

The joint probability distribution is also defined for the causal graph after an intervention [9].

Definition 3.1.2 (Truncated Factorization Due to Do-Intervention)**.** *Given a directed acyclic graph $G = (V, E)$ with n vertices. Let each vertex correspond to a random variable X_j. Suppose*

we intervene, setting $X_j = x'_j$. Then, we can factorize the joint probability distribution as follows,

$$\begin{aligned}\Pr(X_1 =& x_1, X_2 = x_2, \ldots, X_n = x_n | do(X_j = x'_j)) \\ &= \mathbb{1}\{x_j = x'_j\} \prod_{i \in [1,n], i \neq j} \Pr(X_i = x_i | \text{PA}(X_i) = pa(X_i)),\end{aligned}$$

where $\text{PA}(X_i)$ is the set of random variables corresponding to the parents of the node V_i, $pa(X_i)$ are the values of $\text{PA}(X_i)$, and $\mathbb{1}\{x_j = x'_j\}$ takes the value 1 when $x_j = x'_j$ and 0 otherwise.

This definition is very similar to the graph factorization we would get if we just considered the graph after intervention, and factorized it without accounting for the intervention. The only difference from the graph factorization is that instead of the $\Pr(X_j = x_j)$ term, there is a $\mathbb{1}\{x_j = x'_j\}$ term. The $\mathbb{1}\{x_j = x'_j\}$ term captures how an intervention changes the probability distribution for node X_j—since we have set $\text{do}(X_j = x'_j)$, when $x_j = x'_j$ then $\Pr(X_j = x_j)$ is one, and in all other cases it is zero.

Consider the example structural causal model in Figure 3.1(b). We can calculate the probability distribution in the case where we have the do-intervention $X_T = x_t$. In this case, using Definition 3.1.2 the joint probability distribution is,

$$\begin{aligned}&\Pr(X_T = x_t, X_H = x_h, X_A = x_a, X_B = x_b, X_C = x_c, X_D = x_d | do(X_T = x'_t)) \\ &= \mathbb{1}\{x_t = x'_t\} \prod_{i \in \{H,A,B,C,D\}} \Pr(X_i = x_i | \text{PA}(X_i) = \text{pa}(X_i)) \\ &= \mathbb{1}\{t = t'\} \cdot \Pr(X_H = x_h | X_A = x_a, X_B = x_b, X_C = x_c, X_T = x_t) \\ &\quad \cdot \Pr(X_A = x_a) \cdot \Pr(X_B = x_b | X_A = x_a) \cdot \Pr(X_C = x_c | X_T = x_t) \\ &\quad \cdot \Pr(X_D = x_d | X_H = x_h, X_T = x_t).\end{aligned}$$

Notice that the graph factorization above, with the do-intervention, is almost the same as the graph factorization without the intervention. The only change we have made is that instead of the term $\Pr(X_T = x_t | \text{PA}(X_T) = \text{pa}(X_T))$, we now have an indicator function regarding the value of X_T matching the intervention value. Also note that since the only edges we removed were the edges incoming to T, the parent sets of all of the other nodes in the structural causal model have not changed.

We will use this observation to rearrange the two equations in Lemma 3.1.1 and Definition 3.1.2, until we have an equation that relates a probability with a do-intervention to one without.

Let us continue to consider the case where we intervene on X_j setting it to x'_j. We rearrange the equation in Lemma 3.1.1 to resemble the do-intervention equation as follows,

$$\begin{aligned}&\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &\quad = \Pr(X_j = x_j | \text{PA}(X_j) = \text{pa}(X_j)) \cdot \prod_{i \in [1,n], i \neq j} \Pr(X_i = x_i | \text{PA}(X_i) = \text{pa}(X_i)).\end{aligned}$$

We then can isolate the product term to get

$$\frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{\Pr(X_j = x_j | \text{PA}(X_j) = \text{pa}(X_j))} = \prod_{i \in [1,n], i \neq j} \Pr(X_i = x_i | \text{PA}(X_i) = \text{pa}(X_i)).$$

We notice that the right-hand side of this equation also appears in the equation in Definition 3.1.2, so we substitute to get

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | do(X_j = x'_j)) \\ = \mathbb{1}\{x_j = x'_j\} \cdot \frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{\Pr(X_j = x_j | \text{PA}(X_j) = \text{pa}(X_j))}.$$

In the equation above, we are only interested in the case when $x_j = x'_j$. In all other cases, both sides become zero, and this equation is not particularly interesting. Hence, we can substitute $x_j = x'_j$ and remove the indicator variable to get,

$$\frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{\Pr(X_j = x_j | \text{PA}(X_j) = \text{pa}(X_j))} = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | do(X_j = x_j)). \tag{3.1}$$

In our proofs, we will be intervening on a variable T, so we rewrite the equation above with $X_j = T$. We also relabel $X_{j+1}, X_{j+2}, \ldots, X_n$ as $X_j, X_{j+1} \ldots, X_{n-1}$.

Notice that $\Pr(T = t | \text{do}(T = t)) = 1$, since by definition, $T = t$ when we $\text{do}(T = t)$. More generally, it is true that

$$\Pr(X_1 = x_1, \ldots X_{n-1} = x_{n-1}, T = t | \text{do}(T = t)) = \Pr(X_1 = x_1, \ldots, X_{n-1} = x_{n-1} | \text{do}(T = t)).$$

We use this to rewrite the equation 3.1 to become

$$\frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}, T = t)}{\Pr(T = t | \text{PA}(T) = \text{pa}(T))} \\ = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1} | do(T = t)). \tag{3.2}$$

In the Rocq proofs in this chapter, instead of taking the graph factorizations given in Lemma 3.1.1 and Definition 3.1.2 as assumptions, we use equation 3.2. This choice is made to simplify our lemmas in Rocq. In the original definitions, we considered the parents of every node, a concept that would have to be encoded in Rocq. Equation 3.2 only has one set of parents, $\text{PA}(T)$, which we can explicitly define as a subset of $X_1, X_2, \ldots, X_{n-1}$. We also do not need to work with the product operator, which also makes this formulation easier to work with in Rocq.

3.2 Parental Adjustment Formula

We now prove the parental adjustment formula, which is as follows:

Theorem 3.2.1 (Parental Adjustment Formula)**.** *Let* $\mathrm{PA}(X)$ *be the set of parents of* X*, and* $pa(X)$ *be the range of* $\mathrm{PA}(X)$ *and assume* $Y \neq X$*. Then, the causal effect of* X *on* Y *is identifiable and is given by the formula*

$$\Pr(y|do(x)) = \sum_{pa(X)} \Pr(y|x, pa(X)) \Pr(pa(X)).$$

We notice that this theorem is a special case of the backdoor adjustment formula given in Theorem 1.2.5, since the set Z=$\mathrm{PA}(T)$ satisfies the backdoor criterion relative to (T, H).

The first condition of the backdoor criterion is that no node of Z is a descendant of T. Since T is a descendant of each of the elements in $\mathrm{PA}(T)$, none of the variables in $\mathrm{PA}(T)$ can be descendants of T without creating a cycle. As structural causal models are acyclic, we can conclude that all of the variables in $\mathrm{PA}(T)$ are not descendants of T.

The second condition requires that all backdoor paths between T and H are blocked by the set Z. Note that any backdoor path will include some node PA_i, which is a parent of T, with an edge pointing from PA_i to T. We consider the next edge on the backdoor path between T and H. This edge must either point towards or away from PA_i. These two options are depicted in Figure 3.2 (using parent nodes PA_i and PA_j). In the case of the backdoor path through PA_i, we are conditioning on PA_i, which is a mediator, and so we have blocked this backdoor path. In the case of the backdoor path through PA_j, we are conditioning on a confounder, so we have, once again, blocked the backdoor path. Thus, $\mathrm{PA}(T)$ blocks all backdoor paths between T and H.

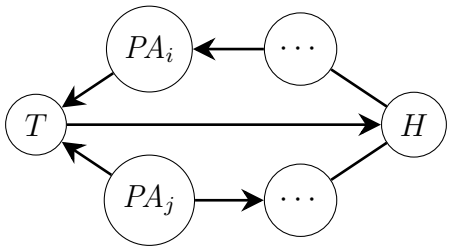


Figure 3.2: Backdoor paths between T and H, all of which include parents

Proving the parental adjustment formula is much easier than the backdoor adjustment formula, and it can be used as a stepping stone to proving the backdoor adjustment formula. To prove the parental adjustment formula we use a different statement of the law of total probability, given in Theorem 2.4.2. Note that by definition, $\Pr(A|B) = \Pr(A, B)/\Pr(B)$ which we use to rewrite the summation term in the law of total probability to get

$$\Pr(A = a) = \sum_{b \in range(B)} \Pr(A = a, B = b). \tag{3.3}$$

We derive the parental adjustment formula by applying the law of total probability to equation 3.2, which we got from factorizing the graph before and after the do-intervention and equating them.

As in Chapter 2, we consider the effect of T on H. First, we rewrite equation 3.2 with the variables T, H and $X_1, X_2, \ldots X_m$, singling out the outcome we are measuring, H, from the rest of the variables, X_i. Rewritten this way the right hand side becomes

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m, H = h | \text{do}(T = t)).$$

In order to compute $\Pr(H = h|\text{do}(T = t))$, we use the law of total probability, as stated in equation 3.3, to get

$$\Pr(H = h|\text{do}(T = t)) = \sum_{(x_1,x_2,\ldots,x_m)} \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m, H = h|\text{do}(T = t)).$$

We perform the same summation on the left-hand side of equation 3.2 to get equation 3.4 as follows,

$$\begin{aligned}\Pr(H = h|\text{do}(T = t)) &= \sum_{(x_1,x_2,\ldots,x_m)} \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m, H = h|\text{do}(T = t)) \\ &= \sum_{(x_1,x_2,\ldots,x_m)} \frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m, H = h, T = t)}{\Pr(T = t|\text{PA}(T) = \text{pa}(T))}.\end{aligned} \tag{3.4}$$

We can partition the variables in $X_1, X_2, \ldots, X_m$ into two sets—those in $\text{PA}(T)$ and a set E containing all the other variables. Note that since neither H nor T can be in $\text{PA}(T)$, $\text{PA}(T)$ must fully be contained in $X_1, X_2, \ldots, X_m$. We split the summation over $(x_1, x_2, \ldots, x_m)$ into two summations—one over $\text{pa}(T)$ and another over e, the possible values of $\text{PA}(T)$ and E respectively. We are not able to simplify the sum over $\text{pa}(T)$ in equation 3.4 since $\text{PA}(T)$ appears in both the numerator and denominator, but we can use the law of total probability to simplify the sum over e to get,

$$\begin{aligned}&\sum_{(x_1,x_2,\ldots,x_m)} \frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m, H = h, T = t)}{\Pr(T = t|\text{PA}(T) = \text{pa}(T))} \\ &= \sum_{\text{pa}(T)} \sum_{e} \frac{\Pr(\text{PA}(T) = \text{pa}(T), E = e, H = h, T = t)}{\Pr(T = t|\text{PA}(T) = \text{pa}(T))} \\ &= \sum_{\text{pa}(T)} \frac{\Pr(\text{PA}(T) = \text{pa}(T), H = h, T = t)}{\Pr(T = t|\text{PA}(T) = \text{pa}(T))}.\end{aligned}$$

At this point, we use the fact that one can rewrite a conditional probability $\Pr(A|B)$ as $\Pr(A, B)/\Pr(B)$. We first expand the conditional probability into a fraction and then collapse it once again but using the other term to achieve a new conditional probability. We use this relation to first expand and then collapse the conditional probability term in the equation above. After these manipulations, we get the right-hand side of the parental adjustment formula as follows,

$$
\begin{aligned}
&= \sum_{\mathrm{pa}(T)} \frac{\Pr(\mathrm{PA}(T) = \mathrm{pa}(T), H = h, T = t)}{\Pr(T = t | \mathrm{PA}(T) = \mathrm{pa}(T))} \\
&= \sum_{\mathrm{pa}(T)} \frac{\Pr(\mathrm{PA}(T) = \mathrm{pa}(T), H = h, T = t) \cdot \Pr(\mathrm{PA}(T) = \mathrm{pa}(T))}{\Pr(T = t, \mathrm{PA}(T) = \mathrm{pa}(T))} \\
&= \sum_{\mathrm{pa}(T)} \Pr(H = h \,|\, \mathrm{PA}(T) = \mathrm{pa}(T), T = t) \cdot \Pr(\mathrm{PA}(T) = \mathrm{pa}(T)).
\end{aligned}
$$

The proof of the parental adjustment formula in Rocq uses the same approach. We begin by defining four random variables, `H`, `T`, `paT` and `E`, as well as their post-intervention versions `Hinterv`, `paTinterv` and `Einterv`. To define these random variables, we define a general sample space and an outcome set for each random variable, and then state the types of the random variables but crucially never define the node functions.

```
Context {R : realType}.
Variables (U : finType).
Variables (outcomesH : finType).
Variables (outcomesT : finType).
Variables (outcomesPaT : finType).
Variables (outcomesE : finType).
Variable P : R.-fdist (U).

Variable H : {RV P -> outcomesH}.
Variable T : {RV P -> outcomesT}.
Variable paT : {RV P -> outcomesPaT}.
Variable E : {RV P -> outcomesE}.

Variable Hinterv : outcomesT -> {RV P -> outcomesH}.
Variable Tinterv : outcomesT -> {RV P -> outcomesT}.
Variable paTinterv : outcomesT -> {RV P -> outcomesPaT}.
Variable Einterv : outcomesT -> {RV P -> outcomesE}.
```

We note that we define E=`E` and $\mathrm{PA}(T)$=`paT` as random variables, without explicitly stating that they are sets of random variables. However, this setup actually allows for them to be lists of random variables, since a list of random variables is still a random variable. We are defining the random variable to map to a generic finite type (`finType`), and since a tuple type of finite types is still a finite type, this definition works regardless of whether `E` is a single finite variable or a list of them. Thus, we declare only these four random variables but are still able to write a lemma about the general case.

We now use the random variables defined above to state our lemma, which matches Theorem 3.2.1,

```
Lemma parental: forall t,
  (forall h pa e,
```

```
    `Pr[ [% (Hinterv t), (paTinterv t), (Einterv t)] = (h, pa, e)]
      = `Pr[ [% H, T, paT, E] = (h, t, pa, e)] / `Pr[ T = t | paT = pa]) ->
  (forall h, `Pr[(Hinterv t) = h] = \sum_(u : outcomesPaT)
      `Pr[ H = h | [% T, paT] = (t, u)] * `Pr[ paT = u]).
```

We formally verify this lemma in Rocq using the proof outline presented above. Note that in this variant of the lemma we do not have explicit assumptions about `paT` being the parents of `T`, or any other graph relations. However, the fact that `paT` is the set of parents of `T` is implicitly recorded in the first assumption, which comes from the graph factorization. When we derived equation 3.2 we relied on the parental relationship between these two variables.

3.3 Backdoor Adjustment Formula from Parental Adjustment Formula

We now move on to proving the backdoor adjustment formula, which we prove by using the parental adjustment formula. We also need a couple of extra assumptions about independence, which can be derived from the structural causal model, in order for this proof to work. Here is the lemma we prove:

Lemma 3.3.1. *If the parental adjustment formula holds,* $\Pr(T = t|\mathrm{PA}(T) = pa(T)) \neq 0$ *for all* $pa(T)$ *and* t*, and*

$$T \perp Z \,|\, \mathrm{PA}(T) \text{ and } H \perp \mathrm{PA}(T) \,|\, T, Z$$

then we can conclude that the backdoor adjustment formula holds with us conditioning on the set Z*. That is,*

$$P(H = h|do(T = t)) = \sum_{Z=z} P(H = h|T = t, Z = z)P(Z = z).$$

We can prove Lemma 3.3.1 by using the law of total probability, and also a new lemma that lets us convert the assumptions about independence of random variables into some statements about probability, as follows

Lemma 3.3.2. *For some random variables* X*,* Y *and* W*, if* $X \perp Y \,|\, W$ *and* $\Pr(Y = y, W = w) \neq 0$ *for some* y *and* w*, then for all* x*,*

$$\Pr(X = x \,|\, W = w) = \Pr(X = x \,|\, W = w, Y = y).$$

Lemma 3.3.2 follows from the definition of independence and already exists in the infotheo library as a lemma.

We can then approach proving the backdoor adjustment formula by taking the parental adjustment formula and introducing a summation over z using the law of total probability. This gets us

$$\begin{aligned}
&\Pr(H = h|\mathrm{do}(T = t)) \\
&= \sum_{\mathrm{pa}(T)} \sum_{z} P(H = h|T = t, \mathrm{PA}(T) = \mathrm{pa}(T), Z = z) \\
&\qquad \cdot \Pr(Z = z|T = t, \mathrm{PA}(T) = \mathrm{pa}(T)) \cdot \Pr(\mathrm{PA}(T) = \mathrm{pa}(T)). \qquad (3.5)
\end{aligned}$$

We then note that using Lemma 3.3.2 and the assumptions in Lemma 3.3.1 we have that

$$P(H = h|T = t, \text{PA}(T) = \text{pa}(T), Z = z) = P(H = h|T = t, Z = z) \quad \text{and}$$

$$\Pr(Z = z|T = t, \text{PA}(T) = \text{pa}(T)) = \Pr(Z = z|\text{PA}(T) = \text{pa}(T)).$$

We can use these two equalities to rewrite (3.5) as

$$\begin{aligned}&\Pr(H = h|\text{do}(T = t))\\ &= \sum_z \sum_{\text{pa}(T)} P(H = h|T = t, Z = z) \cdot \Pr(Z = z|\text{PA}(T) = \text{pa}(T)) \cdot \Pr(\text{PA}(T) = \text{pa}(T)).\end{aligned}$$

We then note that we can now use the law of total probability to sum over $\text{pa}(T)$. Notably, the first term does not depend on $\text{pa}(T)$, and the second two exactly fit the format of the law of total probability. Thus, we get

$$\Pr(H = h|\text{do}(T = t)) = \sum_z P(H = h|T = t, Z = z) \cdot \Pr(Z = z), \tag{3.6}$$

which is exactly the backdoor adjustment formula. Hence, we have proven the backdoor adjustment formula from the parental adjustment formula. We are also able to follow this proof approach in Rocq to formally verify prove the following lemma, which is the same as Lemma 3.3.1.

```
1 Lemma parental_to_cond: forall t,
2   (forall h, ‘Pr[ (Hinterv t) = h] = \sum_(u : outcomesPaT)
3       ‘Pr[ H = h | [% T, paT] = (t, u)] * ‘Pr[ paT = u]) ->
4   T _|_ Z | paT ->
5   H _|_ paT | [% T, Z] ->
6   (forall u t, ‘Pr[ T = t | paT = u] != 0) ->
7   (forall h, ‘Pr[ (Hinterv t) = h] = \sum_(z : outcomesZ)
8       ‘Pr[ H = h | [% T, Z] = (t, z)] * ‘Pr[ Z = z]).
```

In our Rocq lemma statement above, lines two and three state the parental adjustment formula, lines four and five state our independence conditions, line six states the non-negativity condition, and then lines seven and eight state the parental adjustment formula.

We combine Theorem 3.2.1 and Lemma 3.3.1 to get a proof of the backdoor adjustment formula that requires a couple of extra assumptions that differ from those in Theorem 1.2.5. Lemma 3.3.3 states the assumptions we need for our proof.

Lemma 3.3.3. *Consider a structural causal model where we are trying to find the effect that intervening on T causes on H. Let $\text{PA}(T)$ be the set of parents of T, and E be all the other nodes in the graph. If*

1. *for all $h, pa(T), e$, $\Pr(H = h, \text{PA}(T) = pa(T), E = e|do(T = t)) = \Pr(H = h, T = t, \text{PA}(T) = pa(T), E = e)/\Pr(T = t|\text{PA}(T) = pa(T))$,*
2. *$T \perp Z \,|\, \text{PA}(T)$,*

3. $H \perp \mathrm{PA}(T) \,|\, T, Z$,

4. for all $pa(T), t$, $\Pr(T = t | \mathrm{PA}(T) = pa(T)) \neq 0$,

then,

$$\Pr(H = h | do(T = t)) = \sum_z P(H = h | T = t, Z = z) \cdot \Pr(Z = z).$$

We also formally verify this lemma in Rocq, proving

```
1 Lemma graphfactor_indp_backdoor_adj: forall t,
2   (forall h pa e,
3   ‘Pr[ [% (Hinterv t), (paTinterv t), (Einterv t)] = (h, pa, e) ]
4       = ‘Pr[ [% H, T, paT, E] = (h, t, pa, e)] / ‘Pr[T = t | paT = pa]) ->
5   T _|_ Z | paT ->
6   H _|_ paT | [% T, Z] ->
7   (forall u t, ‘Pr[ T = t | paT = u] != 0) ->
8   (forall h, ‘Pr[(Hinterv t) = h] = \sum_(z : outcomesZ)
9       ‘Pr[H = h | [% T, Z] = (t, z)] * ‘Pr[Z = z]).
```

We note that we have a couple of assumptions that remain, none of which appear in the backdoor adjustment formula as stated in Theorem 1.2.5. However, we claim that these assumptions are reasonable and ones that we expect are provable.

Specifically, we note that the first assumption, on lines 2–4, comes from graph factorizations, and we derived it in Section 3.1. Ideally, eventually this assumption would be changed to be assumptions about the two graph factorizations (one for the graph without intervention and one with intervention). However, since we have derived equation 3.2 by hand and it is a fairly short derivation, we believe that we would not need any new assumptions to make this change. Furthermore, we expect that changing this assumption would only require a small amount of work.

The assumptions in lines five and six, about independence, are more interesting and seem like they would be substantially harder to prove. We first argue that these assumptions seem like things one could prove in general. In *Causality* [9], Pearl states Theorem 3.3.4 relating independence to properties of the structural causal model.

Theorem 3.3.4. *If sets* X *and* Y *are d-separated by* Z *in a DAG* G*, then* X *is independent of* Y *conditional on* Z *in every distribution compatible with* G*.*

Thus, when we reason about these independence assumptions, we consider them from the lens of d-separation.

The first assumption is that $T \perp Z \,|\, \mathrm{PA}(T)$. This equates to saying that T and Z are d-separated by $\mathrm{PA}(T)$. Recall that to satisfy the backdoor criterion, the set Z must consist of the descendants of $\mathrm{PA}(T)$. Then, we can conclude that either a path between $Z_i \in Z$ and T goes through a parent of T, or the path contains an edge from T to some node E_k, or $Z_i \in \mathrm{PA}(T)$. We note that in the first case the path is blocked. We can see this in Figure 3.3. After visiting $PA_j \in \mathrm{PA}(T)$, there are two options—either there is an outgoing edge or an incoming edge along the path between T and Z_i. In the first case, this means that PA_j is

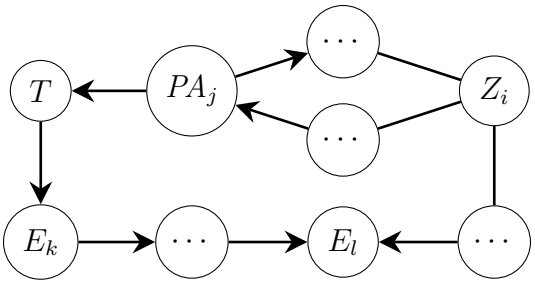


Figure 3.3: Options for backdoor paths between T and Z

a confounder, and in the second PA_j is a mediator, so in either case, conditioning on PA_j blocks this backdoor path.

The other case was where the path between T and Z_i did not immediately visit a parent, but instead went to some node E_k. In this case, we know there is an arrow from T to E_k and not the other way round as E_k is not a parent of T. Since we know Z_i is not a descendant of T, we know that at some point in time the arrows must switch directions. Hence there will be a node, E_l in Figure 3.3, which is a collider along this path. Thus, this path is also blocked.

Then, we must consider when $Z_i \in \text{PA}(T)$. We may have the edge $Z_i \to T$, so it is impossible that these two notes are d-separated. Therefore, we can only conclude d-separation for all nodes $Z_i \in Z \backslash \text{PA}(T)$. Using Theorem 3.3.4, we can then get that $T \perp Z \backslash \text{PA}(T) \,|\, \text{PA}(T)$. To get to the independence condition we were trying to prove we need Lemma 3.3.5.

Lemma 3.3.5. *Given the random variables W, X, Y and Z, if $X \perp Y \,|\, W, Z$, then*

$$X \perp Y, Z \,|\, W, Z.$$

Now, we note that if we conclude $T \perp Z \backslash \text{PA}(T) \,|\, \text{PA}(T)$, we are able to use Lemma 3.3.5 to find that $T \perp Z \,|\, \text{PA}(T)$, which is what we wanted.

We also have formally verified Lemma 3.3.5 in Rocq, proving the following lemma.

```
Lemma adding_conditional_to_indep: forall  {A B C D : finType}
  (X : {RV P -> A}) (Y : {RV P -> B}) (W : {RV P -> C}) (V : {RV P -> D}),
  X _|_ Y | [% W, V] ->
  X _|_ [% Y, V] | [% W, V].
```

We can also consider the second condition, $H \perp \text{PA}(T) \,|\, T, Z$, and similarly reason about d-separation. Once again, we will have to address the case where $\text{PA}(T)$ and Z coincide separately. Beyond that case, there are two possibilities for us to consider—one where we have a path that goes through T, and one were it does not go through T. We can see these two options in Figure 3.4. First, we consider the situation where T has an outgoing edge towards H. With this edge, T is a mediator, so conditioning on T blocks this door path. Next, we consider when there is an edge coming into T as we keep going along path to H. Then, we notice that there is a backdoor path between T and H by going along this path, in which case we know that it must be blocked by Z (since Z satisfies the backdoor criterion with (T, H)), so this path must also be blocked.

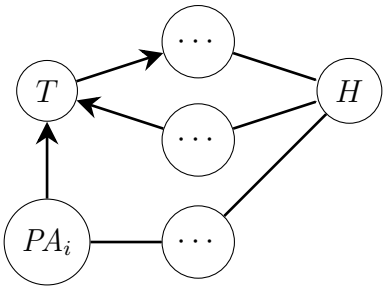


Figure 3.4: Options for backdoor paths between H and $\text{PA}(T)$

Now, we consider the other case where there is a path between PA_i and H, and T is not along this path. This path is a subpath of a backdoor path between T and H, and thus it must be blocked. If this path is blocked by a collider, then the path will still be blocked since PA_i is not a collider (since it has an outgoing edge). If it was blocked by conditioning on a mediator or confounder, then this path is still blocked as long as the mediator/confounder was not PA_i itself.

Now we consider the case where PA_i is the mediator/confounder in Z that blocks this backdoor path between T and H. We note that in this case, we are not able to conclude that the backdoor path is blocked. However, we can conclude that independence holds for all $\text{PA}(T)$ that are not also in Z, so we conclude that $H \perp \text{PA}(T) \backslash Z \,|\, T, Z$. Then, as before, we use Lemma 3.3.5 to conclude that $H \perp \text{PA}(T) \,|\, T, Z$, the independence statement we are interested in, is also true.

3.4 Example Graphs

We have not fully proven the backdoor adjustment formula—in Lemma 3.3.3 we have two independence assumptions that do not exist in the backdoor adjustment formula as stated in Theorem 1.2.5. We consider a smaller four-node graph (shown in Figure 3.5) which we use to verify that the independence conditions are provable in this specific case, as well as double check that none of our assumptions are unreasonable.

We also show how our Rocq lemma works in the general case. We consider an eight-node graph (shown in Figure 3.6) where we do not verify the independence statements but instead show that our lemma works even when the random variables corresponding to $\text{PA}(T), Z$ and E are sets rather than single random variables.

3.4.1 Four-Node Example Graph

We consider the four-node example show in Figure 3.5 and use it to verify that the conditional-independence assumptions we make in Lemma 3.3.2 are satisfiable, at least in this specific example. We also verify that the other assumptions made in Lemma 3.3.2 are valid for this graph. While working with a single sample graph by no means constitutes a proof of any of these assumptions being correct and true in all graphs, working with a sample graph is a useful sanity check to make sure our assumptions are not obviously problematic.

We note that in this graph, if we are looking at the effect of T on H, we have that $\text{PA}(T) = \{C\}$, $E = \{E\}$. For the set Z which satisfies the backdoor criterion with respect to (T, H) we actually have two choices—both C and C, E work. We choose to consider the case where $Z = \{C\}$ since we believe that, on average, in a scientific experiment having smaller conditioning sets is easier to work with. We write the formulation of the backdoor adjustment formula with $Z = \{C\}$ in Lemma 3.4.1.

Lemma 3.4.1. *Given the structural causal model in Figure 3.5,*

$$\Pr(H = h | do(T = t)) = \sum_c \Pr(H = h | T = t, C = c) \cdot \Pr(C = c).$$

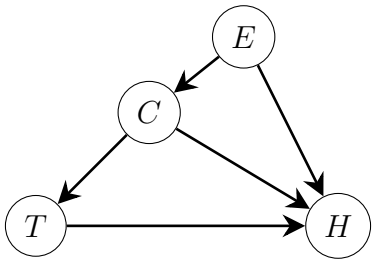


Figure 3.5: A four-node structural causal model with two confounders

It turns out that Lemma 3.4.1 is easy enough to show by working through the parental adjustment formula. We aim to show that starting from the mutual independence of the unobserved factors, we can show conditions two and three of Lemma 3.3.3 hold, and then we are able to show that the backdoor adjustment formula holds in this case with only assumptions regarding graph factorization and non-negativity. We formally verify the backdoor adjustment formula by setting up our node functions in Rocq as we did in Section 2 and then proving the following lemma:

```
1 Lemma four_var_confounder_backdoor_adjustment: forall t,
2   (forall h c e, ‘Pr[ [% (Hinterv t), (Cinterv t), (Einterv t)] = (h, c, e)]
3       = ‘Pr[ [% H, T, C, E] = (h, t, c, e)] / ‘Pr[T = t | C = c]) ->
4   mutual_indep_four UERV UCRV UTRV UHRV ->
5   (forall u t, ‘Pr[T = t | C = u] != 0) ->
6   (forall h, ‘Pr[(Hinterv t) = h] = \sum_(c : outcomesC)
7       ‘Pr[ H = h | [% T, C] = (t, c)] * ‘Pr[C = c]).
```

In this Rocq lemma, lines one and two state the assumption that comes from graph factorization, line three states that the unobserved terms of E, C, T and H are all mutually independent, and line four states the non-negativity constraint. Then lines six and seven actually state the backdoor adjustment formula.

However, we actually notice that since in this example $\text{PA}(T) = Z = \{C\}$, conditions two and three of Lemma 3.3.3, which are the independence conditions, become

$$T \perp C \,|\, C \text{ and } H \perp C \,|\, T, C.$$

These independence statements are actually provable without knowing the mutual independence of the unobserved terms, due to the conditioning on C. Thus, while this example tells us that Lemma 3.3.3 holds in the case of Figure 3.5, it is perhaps not the most interesting example, since the independence statements were provable without using the mutual independence that comes from structural causal models.

3.4.2 Eight-Node Example Graph

We also consider a larger structural causal model with eight nodes, as shown in Figure 3.6. This model is interesting to us because all of the random variables that we claimed could be sets in the previous proof are now actually sets. Specifically, in this structural causal model, we have that $\text{PA}(T) = \{C, D, Q\}$ and $E = \{E, F, G\}$. We also claim that the set $Z = \{C, F\}$ satisfies the backdoor criterion relative to (T, H). We notice that there are only three backdoor paths—$T - D - F - H$ which is blocked by F, and $T - C - E - H$ and $T - C - H$ which are both blocked by C.

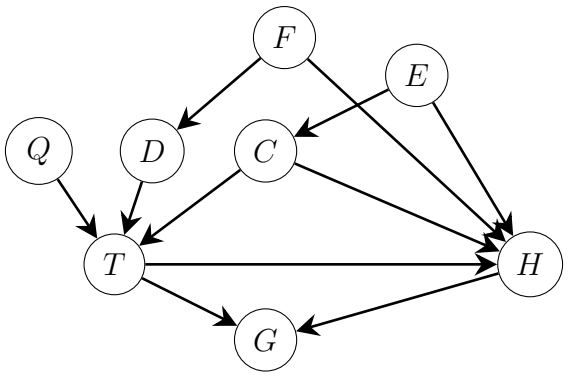


Figure 3.6: An eight-node structural causal model with sets Z, $\text{PA}(T)$ and E of size greater than one

We note that in this case we also have that C is in both $\text{PA}(T)$ and in Z, so these two sets are not disjoint. Thus, we claim that the structural causal model in Figure 3.6, while still being fairly small, is able to encapsulate some of the complexities that come with bigger structural causal models, since it is now working with sets, some of which are partially overlapping.

Nonetheless, we are able to prove the following Rocq lemma for this model:

```
Lemma eight_var_confounder_backdoor_adjustment: forall t,
  (forall h pa e,
  ‘Pr[ [% (Hinterv t), [% (Cinterv t), (Dinterv t), (Qinterv t)],
      [% (Einterv t), (Finterv t) ]] = (h, pa, e) ]
      = ‘Pr[ [% H, T, [% C, D, Q] , [% E, F]] = (h, t, pa , e ) ]
        / ‘Pr[ T = t | [% C, D, Q] = pa ]) ->
  T _|_ [% C, F] | [% C, D, Q] ->
  H _|_ [% C, D, Q] | [% T, [% C, F]] ->
```

```
9    (forall pa t, ‘Pr[ T = t | [% C, D, Q] = pa ] != 0) ->
10   (forall h, ‘Pr[(Hinterv t) = h] = \sum_(cf : outcomesC * outcomesF)
11       ‘Pr[ H = h | [% T, [% C, F] ] = (t, cf)] * ‘Pr[[% C, F] = cf]).
```

This lemma makes the same assumptions as were made in Lemma 3.3.3. However, they are all modified to be specific to this case. Our assumptions are as follows:

1. Lines 2–6 assume the equality we get from rearranging the graph factorization formulas, which we also saw in Lemma 3.3.3, but specific to our nodes (H, T, C, D, Q, E, F).

2. Line seven contains our independence statement $T \perp Z \,|\, \mathrm{PA}(T)$, but we substitute $Z = \{C, F\}$ and $\mathrm{PA}(T) = \{C, D, Q\}$.

3. Line eight contains our independence statement $H \perp \mathrm{PA}(T) \,|\, T, Z$, where we once again substitute $Z = \{C, F\}$ and $\mathrm{PA}(T) = \{C, D, Q\}$.

4. Line nine contains the assumption that for all $\mathrm{pa}(T), t$, $\Pr(T = t | \mathrm{PA}(T) = \mathrm{pa}(T)) \neq 0$, but once again $\mathrm{PA}(T) = \{C, D, Q\}$.

In the final two lines, we are able to conclude that the backdoor criterion holds in this case, finding

$$\Pr(H = h | \mathrm{do}(T = t)) = \sum_{(c,f)} P(H = h | T = t, C = c, F = f) \cdot \Pr(C = c, F = f).$$

Note that we never prove that, for this causal model, the independence assumptions hold, so we do not have a complete proof of the backdoor adjustment formula for this case. Rather, this example demonstrates that our formally verified proof of Lemma 3.3.3 is actually capable of working with sets of random variables for $\mathrm{PA}(T), E$ and Z and is also capable of handling overlapping elements between $\mathrm{PA}(T)$ and Z, and between E and Z, at least for one specific example.

3.5 Implementation Details

We note the Rocq axioms that we used in order to prove the lemmas in this chapter. These axioms are the same as the ones we used for Chapter 2. Rocq is built on the Calculus of Inductive Constructions [15], so we have used the inference rules associated with it. We also use the axioms

- `boolp.propositional_extensionality`,
- `boolp.functional_extensionality_dep` and
- `boolp.constructive_indefinite_description`.

Beyond these axioms, all of our proofs were fully proven, and we did not have to make any more assumptions besides those explicitly stated in the lemmas.

Chapter 4

Application to Scientific Experiments

In this chapter, we will look at some papers that collect data and then run regressions to find correlations between their variables. We argue that they could have instead constructed causal diagrams in order to understand the causal effects rather than just the correlations. To discuss the causal effects, we construct a structural causal model on the variables they consider in each paper. We also show how the backdoor adjustment formula could have been applied with the data they have collected, formally verifying the formulas we present using Rocq. We argue that using the backdoor adjustment formula would have yielded a slightly more interesting result than the one they report in the paper.

4.1 Electronic Media and GPA

In this section, we look at the paper "The Wired Generation: Academic and Social Outcomes of Electronic Media Use Among University Students" [16]. This paper, published in 2011, was one of the earlier papers trying to explore the effect of electronic media use on academic outcomes.

4.1.1 Paper Results

To study the effect of electronic media use on GPA the authors first consider "Model 1," where they account for electronic media use (splitting into separate types of electronic media, such as social networking sites, calling and texting, email, and video games) and other time use (academic, social face-to-face, or other). They run a regression on all of these factors and then note the ones with strong predictive power. The regression shows that electronic media time-use accounts for 6% of the variance in GPA.

Then, the authors consider "Model 2" where they control for some more variables. Specifically, they control for:

- demographics (gender, age, martial status, international student, on-campus or off-campus housing),
- number of credits taken that semester,
- whether the student is a transfer student,

- ACT score, and
- employment.

They run another regression allowing these variables to also impact the final GPA score. Here electronic media time-use accounts for 22% of the variance in GPA. Hence, the authors conclude that electronic media use is negatively associated with grades.

4.1.2 Causal Interpretation

We claim that the question of interest is not just whether there is an association between electronic-media and grades—it is more interesting to know whether electronic media use causes lower grades. It is not impossible that there is some other factor, such as worse self-discipline, that could cause students both to spend more time on electronic media and to have lower grades. In such circumstances, reducing the time spent on electronic media would not necessarily improve grades, since the underlying self-discipline problem which is causing the lower grades persists.

Additionally, we note that as we move to a causal framework we can now also model non-linear relationships, which is something that could not be captured by the regression that was used in the original paper.

By controlling for some variables, the authors have identified the factors that they believe are likely to have causal effects on a student's GPA. We take these variables and draw a causal model where each of them has an effect on GPA. We also add causal connections where it seems likely that there is a causal effect—for example, we connect demographics (D) to how they use their time, besides spending it on electronic media (O). We believe that information captured in the student demographics, such as whether they are married or not, may affect how they spend their time. This causal diagram is shown in Figure 4.1.

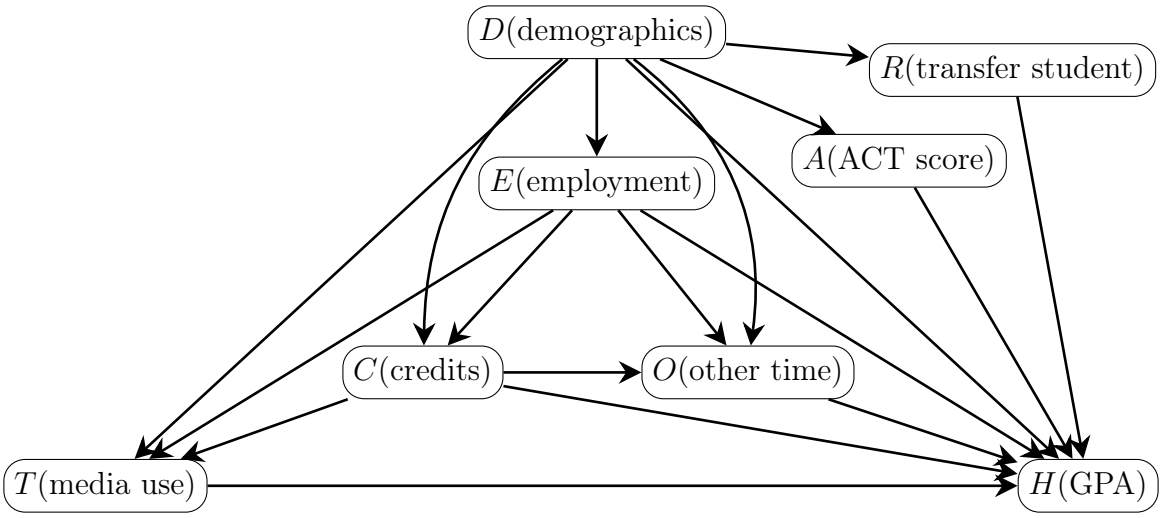


Figure 4.1: Causal model showing variables that affect electronic media use and GPA, as well as causal relationships between them

We can then use the backdoor adjustment formula to find an equation for $\Pr(H|\text{do}(T = \text{none}))$. Note that $\Pr(H|\text{do}(T = \text{none}))$ corresponds to intervening to set electronic media use to none—in other words, this intervention is a ban on electronics. This is a quantity that we are, arguably, interested in, since schools are able to (and many have) adopt policies to ban electronics.

In this case we notice that a set Z that satisfies the backdoor criterion with respect to (T, H) is $Z = \{C, D, E\}$. In fact, we argue that we must include C, D and E in our set since each of them is a confounder on a three-node backdoor path. For example, we have the path $T - D - H$. The only way to block that backdoor path is by including D in Z. It so happens that this set Z also blocks all other backdoor paths.

We notice that the set $\{C, D, E\}$ is also the set of parents of T, so we can use the parental adjustment formula and apply it to this case. We define our node functions based on the causal network shown in Figure 4.1. We represent each node function by the letter shown in the figure. So, for example, we have that $C = f_C(U_C, D, E)$, where U_C is the unobserved term for the random variable C. Then we use Rocq to formally verify the backdoor adjustment formula in this case. Since this is the parental adjustment formula, we can rely on the Lemma proven in Chapter 3, making this proof very short, and giving us Lemma 4.1.1 with effectively no further work on our part.

Lemma 4.1.1. *Consider the structural causal model shown in Figure 4.1. If the equality derived from the graph factorizations,*

$$\begin{aligned}\forall h, c, d, e, a, n, o.&\\ &\Pr(H_{T=0} = h, C_{T=0} = c, D_{T=0} = d, E_{T=0} = e, A_{T=0} = a, N_{T=0} = n, O_{T=0} = o)\\ &= \frac{\Pr(H = h, T = t, C = c, D = d, E = e, A = a, N = n, O = o)}{\Pr(T = t|C = c, D = d, E = e},\end{aligned}$$

holds and $\Pr(T = 0|C = c, D = d, E = e) \neq 0$*, then*

$$\begin{aligned}\forall h.\quad &\Pr(H_{T=0} = h)\\ &= \sum_{(c,d,e)} \Pr(H = h|T = 0, C = c, D = d, E = e) \cdot \Pr(C = c, D = d, E = e).\end{aligned}$$

Since this data was collected via a survey of students, it should be possible to calculate each of the observational probabilities in Lemma 4.1.1 from the data. Thus, from the raw data collected from this experiment one can report the effects of interventions, rather than only reporting the correlations as were reported in this paper. We also note that while regressions generally work with continuous variables, in this experiment all of their variables are actually discrete. Our Rocq proofs assumed that the random variables were discrete, so is applicable in this case. In fact, in many experiments the random variables are discrete, since it is often somewhat hard to measure variables continuously, and many variables are naturally discrete.

4.2 Alcohol Advertisement and Alcohol Consumption

In this section, we look at another paper, "Alcohol Advertising and Young People's Drinking" [17]. This paper studies the effects of alcohol advertisements on alcohol consumption in 17–21 year-old secondary school and university students in Leicester, United Kingdom.

4.2.1 Paper Results

This paper attempted to answer a number of research questions related to alcohol advertisements and their effects on young people. We focus on their second research question: Does reported exposure to alcohol advertising predict frequency of alcohol consumption in the presence of controls for family and peer-group variables?

To answer this question, there was a self-reported questionnaire that participants filled out. The questionnaire asked questions about a number of variables, including:

- alcohol consumption,
- exposure to alcohol advertisements,
- media consumption (amount of TV viewing; amount of TV viewing after 9pm, when alcohol advertisements are move prevalent),
- demographics (gender; age),
- parental attitudes towards alcohol (how much their parents drink; whether their parents have offered them alcohol; parents' reaction to them getting drunk), and
- friends' attitudes towards alcohol (how many friends drink; how often friends meet in locations where alcohol is served).

The researchers then ran a regression on the variables listed above, to find that alcohol advertising did not have a substantial impact on alcohol consumption, but parental and peer-group attitudes did.

4.2.2 Causal Interpretation

We, once again, consider this study from a causal lens. We add the variables that were used in the regression into a causal graph as well as edges where it seems likely that there would be causal effects. This causal graph is shown in Figure 4.2.

We note some differences from the experiment in Section 4.1. Firstly, they considered media consumption in their regression—the researchers believed that media consumption could affect alcohol consumption. However, it does not seem very plausible that media consumption, which in this case is how much TV the subjects watch, directly affects their alcohol consumption. It may affect their alcohol consumption due to the fact that one may see more alcohol advertisements, but this effect is already represented on the causal graph by the path $M \to T \to H$.

We also note that F is a mediator between T and H in Figure 4.2. We added an edge from T to F since it is plausible that the alcohol advertisements seen by subjects can affect their friend groups' attitudes towards alcohol. For example, one of the questions about friends' attitudes towards alcohol included asking how often the subject met in locations where alcohol is served. Since the subject is part of the friend group and gets to have a say in where their friends meet, if the subject sees more alcohol advertisements then they may be more willing to meet in such locations.

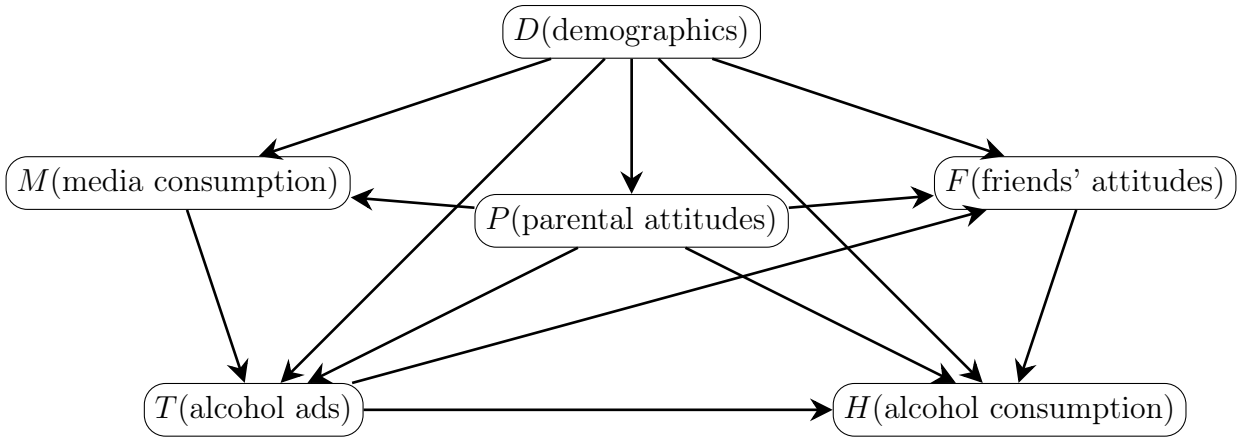


Figure 4.2: Causal model showing the effect of alcohol advertisements on alcohol consumption

Adding the edge $T \to F$ creates a path $T - F - H$, where F is a mediator between T and H. The researchers ran a regression where F was one of the variables. Controlling for F may have caused them to lose some of the effect that T has on H, since any effect that T has on F and then F subsequently has on H would be lost.

When we consider the causal model and apply the backdoor criterion, we can avoid such pitfalls. In this case, we claim that an interesting question, which could be answered with the data collected, is what might happen if there were regulations limiting alcohol advertisements. This corresponds to the probability,

$$\Pr(\text{alcohol consumption} \,|\, \text{do}(\text{alcohol ads} = \text{less})).$$

We note that in this case the set $Z = \{D, M, P\}$ satisfies the backdoor criterion relative to (T, H). Just as in Section 4.1, this set is exactly the parents of T, so we can apply the parental adjustment formula, which we have already verified in Rocq. With minimal work, we formally verify the following lemma in Rocq:

Lemma 4.2.1. *Consider the structural causal model shown in Figure 4.2. If the equality derived from the graph factorizations,*

$$\begin{aligned}\forall h, d, f, m, p. \quad & \Pr(H_{T=less} = h, D_{T=less} = d, F_{T=less} = f, M_{T=less} = m, P_{T=less} = p) \\ & = \frac{\Pr(H = h, T = t, D = d, F = f, M = m, P = p)}{\Pr(T = t | D = d, M = m, P = p)},\end{aligned}$$

holds and $\Pr(T = less | D = d, M = m, P = p) \neq 0$*, then*

$$\begin{aligned}\forall h. \quad & \Pr(H_{T=less} = h) \\ & = \sum_{(d,m,p)} \Pr(H = h | T = less, D = d, M = m, P = p) \cdot \Pr(D = d, M = m, P = p).\end{aligned}$$

In this case, however, we can take a set Z that is smaller than the parents of T. Specifically, the set $Z = \{D, P\}$ also satisfies the backdoor criterion relative to (T, H). We use Rocq to

verify that the backdoor adjustment formula would hold with the set $Z = \{D, P\}$. Since $Z \neq \text{PA}(T)$, we cannot use the parental adjustment formula. Thus, we use the backdoor adjustment formula stated in Lemma 3.3.3, and begin by proving the two independence conditions. We get the following lemma, which we have also formally verified in Rocq:

Lemma 4.2.2. *Consider the structural causal model shown in Figure 4.2. If the equality derived from the graph factorizations,*

$$\begin{aligned}\forall h, d, f, m, p. \quad & \Pr(H_{T=less} = h, D_{T=less} = d, F_{T=less} = f, M_{T=less} = m, P_{T=less} = p) \\ & = \frac{\Pr(H = h, T = t, D = d, F = f, M = m, P = p)}{\Pr(T = t | D = d, M = m, P = p},\end{aligned}$$

holds, $\Pr(T = less | D = d, M = m, P = p) \neq 0$*, and the unobserved terms associated with each variable are independent, then*

$$\begin{aligned}\forall h. \quad & \Pr(H_{T=less} = h) \\ & = \sum_{(d,p)} \Pr(H = h | T = less, D = d, P = p) \cdot \Pr(D = d, P = p).\end{aligned}$$

Similarly to the experiment described in Section 4.1, the data was collected in a survey, such that all of the probabilities listed in the equations given by Lemmas 4.2.1 and 4.2.2 can be calculated from the data gathered. Furthermore, the data collected was actually discrete. There are questions that theoretically could have continuous answers, such as the amount of alcohol consumed. However, to simplify the survey, responses were collected on either a four-point or a seven-point scale. So, once again, our assumption that we are working with discrete random variables does not cause us any problems.

In this case, having a smaller set Z is not actually that helpful, since the data has already been collected and contains information about media consumption. However, if the media consumption data were incomplete, we would be able to use Lemma 4.2.2 but not Lemma 4.2.1. Similarly, if we were to draw the causal diagram before collecting the data, we could decide not to collect data regarding media consumption, since we would know that we could use Lemma 4.2.2.

By using the backdoor adjustment formula, we are able to estimate the effect of an intervention—a policy requiring less alcohol advertisements. We also are able to work with non-linear relationships, unlike a regression. We also handle the case of a potential mediator, the friends' attitudes towards alcohol, without obscuring part of the effect of alcohol advertisements on alcohol consumption.

Chapter 5

Future Work

This thesis proves the backdoor adjustment formula for specific structural causal models. Then, it provides a proof of the backdoor adjustment formula that requires some extra assumptions about our graph. There is still future work that can be done to finish the proof of the backdoor adjustment formula. There are also more relationships between interventional and observational probabilities that can be explored. Finally, there is a lot more things to be done to make the work done here useful as a tool for designing scientific experiments.

5.1 Backdoor Adjustment Formula Proof

The main area where further work is necessary is the proof of the backdoor adjustment formula. We are still missing a large chunk of the proof, showing us why the two independence conditions are true. Filling in this gap is necessary to have a fully verified proof of the backdoor adjustment formula.

To prove that the independence conditions hold, we believe one would need to prove Theorem 3.3.4, which says that one can convert a statement about d-separation into one about independence. We trust that this theorem is correct as it is stated in [9] as well as many other locations, and there are classical proofs of this theorem. However, we think that formalizing this proof will take non-negligible work since this theorem connect a graphical concept, d-separation, with a probabilistic concept, independence.

Additionally, we also need to prove that the d-separation conditions hold. We have sketched out classical proofs as to why we believe them to be true in Section 3.3, and we believe that these outlines could be used to formally verify these conditions. However, there is currently no formal verification of this d-separation, which is also necessary to complete a formally verified proof of the backdoor adjustment formula.

We also recall that the first assumption we had in Lemma 3.3.3 was about graph factorizations; however, it was not stated explicitly in terms of graph factorizations of the graph with and without intervention, but rather as an equality between the factorizations one gets from the two of them. This was for ease of notation, but for completeness sake, a final version of a proof of the backdoor adjustment formula would use the graph factorizations before any rearrangement. However, unlike the independence assumptions, we believe this change would not be very difficult to implement, since there is not very much rearranging done to get to

the equation we use.

5.2 Exploring Assumptions in Theorems

We could also explore whether all of the non-negativity constraints that are stated in the theorems are truly necessary. In general, when writing these proofs, we were willing to add non-negativity constraints whenever it was convenient, as long as the constraints did not seem unreasonable. In some cases, these constraints could perhaps have been avoided or weakened, especially given that the final formulas involve summation and zero terms disappear when summed. While we believe that all of our non-negativity constraints are reasonable and will not cause problems for most experiment designs, it would be better if we were certain that all such constraints are in the least-demanding form possible.

5.3 Connecting to the Semantic Meaning of a Structural Causal Model

Currently, all of our proofs are entirely probabilistic, where we encode the structural causal model as a set of random variables, and all the dependencies that are visible in the graph are encoded as dependencies between the random variables. Eventually, it would make sense to be able to start with a graph and then derive the random variables associated with each node from there, arriving at the theorems proven in this thesis. Work done by Anna Zhang [8] defined structural causal models, both syntactically and semantically. We believe that connecting to Zhang's work would be a good way to incorporate the graphical concepts. Additionally, due the definitions of d-separation, parents, and many other such relevant concepts, we think that using Zhang's framework would be useful in completing the proof of the backdoor adjustment formula. These definitions would be necessary in order to be able to conclude some of the d-separation properties we need for Lemma 3.3.3.

5.4 Interventional and Observational Relationships Beyond the Backdoor Adjustment Formula

We note that there are a number of other useful relationships between interventional and observations probabilities. We have started to prove the backdoor adjustment formula, but there are more formulas that can be used when the backdoor criterion is not satisfied. For example, one could try to prove the front-door adjustment formula. Beyond that, the second rule of do-calculus also relates iterventional and observational probabilities. Formally verifying all three rules of do-calculus would allow us to engage fully with the do-operator and work with it in situations beyond those described in the various adjustment formulas.

5.5 Formally Verifying Experimental-Design Validity

This work is part of a larger effort in using Rocq to create a formally verified framework for experiment design. While this work focuses on verifying the backdoor adjustment formula, there are many more steps required to make something that is useful to scientists when designing experiments. To start with, there are many more theorems and proofs that need to be verified. Beyond that, one needs to design a way for scientists to be able to interface with all of the theorems that are proven through these proofs.

We believe that continuing this works in such a direction would be very useful. Scientific experiments are constantly conducted, and results are used to inform decisions that can have a large impact on our wold. That said, not infrequently, scientific experiments also fail to be reproducible or interpretable. We believe that putting together a formally verified framework that relies on structural causal models can address this issue.

We choose to look at structural causal models since they encode the causal relationships between variables, the very thing scientific experiments try to uncover. By creating a formally verified backbone that allows us to write and verify theorems regarding causal relationships, we help scientists to harness the power of causal reasoning. Work done by Zhang [8] already began this process by defining the concept of structural causal models in Rocq. This work moves further in this direction by verifying the backdoor adjustment formula, a useful formula for understanding relationships between interventional and observational probabilities.

There is other work being done to formalize experiment designs and come up with frameworks that scientists may use when considering an experiment. For example, PLanet [11] is a recent language that formalizes experiment designs, and the paper suggests possible assignments for test subjects, subject to the experiment's constraints. This work, however, does not use formal-verification, and we see great potential in formally verifying claims made in frameworks such as PLanet. We believe that building formally verified frameworks for experiment design could greatly improve experiment validity, interpretability and reproducibility.

References

[1] A. Deaton and N. Cartwright. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine*, **210**, 2018. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue, pp. 2–21. ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2017.12.005.

[2] J. P. Vandenbroucke. "The HRT controversy: observational studies and RCTs fall in line." *The Lancet*, **373**(9671), 2009, pp. 1233–1235. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(09)60708-X.

[3] The ROCQ Development Team. *The Rocq Proof Assistant*. Version 9.2.0. 2025. URL: https://rocq-prover.org/.

[4] T. Haavelmo. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica*, **11**(1), 1943, pp. 1–12. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1905714 (visited on 04/26/2026).

[5] J. Pearl. "Trygve Haavelmo and the Emergence of Causal Calculus." *Econometric Theory*, **31**(1), 2015, pp. 152–179. DOI: 10.1017/S0266466614000231.

[6] J. Pearl. "[Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention." *Statistical Science*, **8**(3), 1993, pp. 266–269. DOI: 10.1214/ss/1177010894.

[7] J. Pearl. "Causal diagrams for empirical research." *Biometrika*, **82**(4), Dec. 1995, pp. 669–688. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.669.

[8] A. Zhang. "Formalizing Causal Models Through the Semantics of Conditional Independence." Available at http://adam.chlipala.net/theses/azhang03.pdf. Master's thesis. Cambridge, MA: Massachusetts Institute of Technology, May 2025.

[9] J. Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009.

[10] J. A. List, S. Sadoff, and M. Wagner. *So you want to run an experiment, now what? Some Simple Rules of Thumb for Optimal Experimental Design*. Working Paper 15701. National Bureau of Economic Research, Jan. 2010. DOI: 10.3386/w15701.

[11] L. Bielicke, A. Zhang, S. Tyagi, E. Berger, A. Chlipala, and E. Jun. *PLanet: Formalizing Experimental Design*. May 2025. DOI: 10.48550/arXiv.2505.09094.

[12] S. Levitt and J. List. "What Do Laboratory Experiments Tell Us About the Real World?" *Journal of Economic Perspectives*, **21**, Jan. 2005.

[13] E. Bareinboim, J. D. Correa, and C. Jeong. *Causal Fusion*. URL: https://causalfusion.net/app.

[14] R. Affeldt, M. Hagiwara, J. Senizergues, J. Garrigue, K. Sakaguchi, T. Asai, T. Saikawa, N. Obata, and A. Bruni. *Formalization of Information Theory*. Aug. 2012. URL: https://staff.aist.go.jp/reynald.affeldt/shannon/.

[15] *Calculus of Inductive Constructions – Coq 8.9.1 documentation*. https://rocq-prover.org/doc/{V}8.9.1/refman/language/cic.html. [Accessed 26-04-2026].

[16] W. C. Jacobsen and R. Forste. "The Wired Generation: Academic and Social Outcomes of Electronic Media Use Among University Students." *Cyberpsychology, Behavior, and Social Networking*, **14**(5), 2011, pp. 275–280. DOI: 10.1089/cyber.2010.0135.

[17] B. Gunter, A. Hansen, and M. Touri. "Alcohol advertising and young people's drinking." *Young Consumers*, **10**, Mar. 2009. DOI: 10.1108/17473610910940756.