rijksuniversiteit groningen
faculteit wiskunde en natuurwetenschappen
faculteit wijsbegeerte

Probabilistic and counterfactual theories of causation

Bachelorondersoek

Wiskunde en

Wijsbegeerte van een bepaald Wetenschapsgebied

Februari 2012

Student: Mirjam A. de Vos

Begeleider Wiskunde: prof. dr. E. C. Wit

Begeleider Wijsbegeerte: prof. dr. B.P. Kooi

Contents

1 Introduction

In this thesis, I want to give an overview of a number of theories of causality. First, I will give some general information about causation and about problems there are with defining it (Section 2). Then I give some special attention to two possible theories of causation, namely the probabilistic theory (Section 3) and the counterfactual theory (Section 4). Besides that I work out some problems regarding confounding (Section 5). Sections 2 and 4 are mostly philosophical, while Sections 3 and 5 are more mathematical, although for reading the section about the counterfactual theory some mathematical way of thinking can be helpful to understand it.

The section about the meaning and problems of causation starts with an example that will be used in the rest of the text. I will first talk about what is meant by causality and what our intuitions say about it, which is followed by a categorization of different theories of causality. After that you can read about a variety of problems that arise when defining causation.

In the section on probabilities of causal relations is explained what directed acyclic graphs are and how things like d-separation work, so we can make causal Bayesian networks. This all in part of the probabilistic theory of causation that people like Pearl [2009] use to deal with relations between events that might be causal.

In Section 4, we will look at some different kinds of implication that can be used in logic to draw conclusions about the truth value of sentences. Besides that, we will have a short look at the counterfactual method that the researchers of the example introduced in Section 2 used in their study.

The last section is about the problem of confounding. We will look at some difficulties of causal inference and at an example that might be confounded by extraneous factors. We become acquainted with the proportional hazard models that are used in this example and with some problems that can arise when one works with these models.

In this thesis, I use the words 'causation' and 'causality' as if they mean the same. Some people do not agree with that [Hulswit, 2002], but here no difference is taken into account.

At the end of this introduction, I would like to thank my supervisors Ernst Wit and Barteld Kooi for the good comments they gave when they had read one of my concept theses.

2 Causation: meaning and problems

In this section we introduce a running example, which will show the characteristics of causality and illustrate the problems that arise when defining it. Causality can be viewed in many different ways, therefore there are many different theories about it. What I want to do, is look at the relationship between the described theories and my intuitions about causation. Besides that, I will ask myself whether it is a bad thing when these two things appear to be very different.

The example to which some of the theories will be applied, is from an article that was published in *Criminology* by Sampson et al. [2006]. Their claim is that marriage causally inhibits crime and they give a counterfactual approach to within-individual causal effects. In Section 4 this approach is explained.

The data used for this study are from Glueck and Glueck [1950, 1968], who followed 500 men from adolescence (age 10-17) to age 32. Sampson and Laub made sets of additional data, with the same subjects of the former study. They followed the men that were still alive and traceable to age 70 and had in-depth interviews with 52 of them. In the article, the authors state this is the longest longitudinal study to date on crime and adult development. They come to the following conclusion: "We found that being married is associated with a significant reduction in the probability of crime, averaging approximately 35 percent across key models in both the full sample of nearly 500 men examined from ages 17 to 32 and the targeted subsample of 52 men assessed from ages 17 to 70." [Sampson et al., 2006, p. 498]

Using this example, some things about causality will be made clear, including the problems there are in defining it. Although the study does not see crime-rate as a binary variable, we will look at it in that way, because that will make the examples in the theories less complicated.

2.1 What does causality mean?

In general, causality is a relation between events. The structure of a causal claim is 'C causes E', where the events C (cause) and E (effect) are called *relata*. In this subsection you can read about some different theories and intuitions about causation and about the difference between causation and correlation.

2.1.1 Many different theories

The goal of a theory of causality is make a distinction between causal and accidental relations between events. There are many different ways to look at the causal relation between events. One is the way in which it is done in regularity theories, where the definition of a cause is something like 'an event that is always followed by the same other event'. For example, the first event is 'a stone is thrown in the direction of a window' and the second event is 'the window at which a stone was thrown breaks'. So for these theories, a relation can be named causal if and only if it is a relation for all equal kind of events.

Another approach is the probabilistic one, in which the central idea is that a cause changes the probability of its effect [Hitchcock, 2011]. For example if it is known that the chance of committing any crime is lower for a criminal when he is married than when he is not, the conclusion could be that for a criminal being married causes not committing crime (anymore).

Another approach to look at causal relations is the counterfactual way. The idea of counterfactual theories of causation is that the meaning of causal claims can be explained in terms of counterfactual conditionals of the form 'If C had not occurred, E would not have occurred' [Beebee et al., 2009, Chapter 8]. For example the claim that for a criminal being married causes committing no crime, can be explained in terms of the counterfactual 'If *this man is married* had not occurred, *this man commits no crime* would not have occurred' or 'If this man was not married, he would still commit crime'. These last two kinds of theories can be combined in probabilistic counterfactual theories, that deal with sentences like 'if C had not occurred, E would not have had the probability of occurring that it did have'.

There are many other different kinds of theories of causation, such as, causal process theories and agency and interventionist theories. Within these views of causal relations, there are again many different theories, developed by different people. We can categorize these theories using the answers they give on some key questions, as we will see in Section 2.2.

2.1.2 Intuition about causality

To compare theories and our intuition about causality, we will first describe what we understand by that intuition.

If it is said that turning a light switch causes the current to flow or halt, it is about a specific situation, but in some way it seems to mean that turning any light switch at any moment would cause the current to flow or halt.

This is an example that is discussed in Iwasaki and Simon [1986]. If C is a cause of E, where for example C is 'turning this light switch now' and E is 'a current is flowing in this circuit', it is expected that every event like C will cause an event like E. In the next sections we will see whether the theories that will be discussed can agree with the idea that 'C causes E' means 'all events like C cause an event like E'. Another way of looking at this, is that there can always be exceptions on causal relations. For example, if smoking causes cancer, there still might be people that smoke, but never get cancer. So in this case, different people can have different intuitions. Some theories will agree with the former, that says that two relata are causally related if and only if between all events similar to the relata have the same relation. Other theories will agree with the latter, that accepts exceptions on this idea.

If you accepts the first mentioned intuition, there are always necessary and sufficient conditions which determine whether the relation can be regarded as causal [Pearl, 2009, Chapter 9]. This distinction is important in order to prevent incorrect causal statements. The claim 'C causes E' is surely correct if the sufficient condition is satisfied, but not necessarily true if only the necessary condition is satisfied. If C is defined as 'x gets married', then the necessary condition for 'C causes E' being true is that for all x in our domain (all investigated persons), E is the case (E is for example 'x commits no crime'). If the sufficient condition is satisfied, there is indeed causation (although this might be a false inference of causation because of other factors, but we will come to that later). The necessary condition in this case is that for all x for which E is the case, also C is the case. If the necessary condition is not satisfied, the claim is surely false.

Besides that, our intuition (at least my intuition and the intuition of most of my fellow students) says that it is impossible that an effect of a cause appears before the cause itself. So if a theory of causation does not forbid that possibility, it does not correspond to our intuition. Something that might be possible, is that an event A and an event B are causing each other (at the same time). For example if two people are studying together for a test, it might be the case that they stimulate each other (if one stops studying, the other might be less motivated to continue). From the moment they both are studying, it can be the case that 'person X is studying hard' and 'person Y is studying hard' are causing each other. We call this kind of causal relation a *causal circle*.

In addition to this, there has to be a clear distinction between the cause and the effect in cases where we do not assume there is a causal circle. A theory that does not make that difference, makes it impossible to determine whether 'C causes E' is true or 'E causes C'. It is important that this can be

done, because the sentences have different meaning.

So we have three intuitive ideas about causality: C causes E if and only if every event like C causes an event like E (with or without exceptions, in the second case including the necessary and sufficient conditions); a cause is an event that happens before the effect of it; the cause and the effect in a causal relation can not exchange except if there is a causal circle. All these things are intuitions about causality that will be compared with the ideas of the theories that are to be discussed. It is possible that these ideas are very different from the described intuitions. If that is the case, we can conclude that our theories are false, or that our intuitions are not correct, or that both the theories and the intuitions are not right. If for example a theory makes no distinction between the cause and the effect in a causal relation, so $A \rightarrow B$ means the same as $B \rightarrow A$, the theory might not be as useful as we want. There might be a case in which we expect a causal relation, but in which we have no idea whether one event is the cause or the effect, or if it might be that there is a causal circle. In that case, we want the theory to tell us which one of that is true. If the theory makes no distinction between cause and effect, it can not tell us what we want to know, so the theory fails at least partially.

In the latter case, we assumed that if our intuition tells us that there are causes that are no effects (and effects that are no causes), it is indeed the case that such things exist. If we then have a theory that does not agree with that intuition, it might also be the case that the theory is true, but our intuitions are not. Then we should probably try to change our intuitions so they match with the theory. It depends on the amount at which we are convinced of the rightness of the theory if we do that.

2.1.3 Causation versus correlation

It is important to realize that causation is not the same as correlation, since if we ignore this distinction we would easily draw wrong conclusions about the causality of the relation between two events.

If two events are correlated, it seems logical to conclude that one of them causes the other one. If someone looks at a group of non-married criminals and notices after a number of years that the ones that do marry commit less crime per person than the ones that stay single, it is simple to conclude that marriage causes committing less crime. This might be a wrong conclusion.

According to Oxford dictionaries [Stevenson, 2010], a correlation is 'a mutual relationship or connection between two or more things'. For example the event that I am thinking about causation and the event that I am

typing this section are correlated, because they are mutual related because they take place at the same time. Is one of these events causing the other one, or could it be that the events are correlated without being causal related? We will check whether the two possible causal relations, namely the option that me thinking about causation causes me typing this and the option that me typing this causes me thinking about causation, are necessary true.

Suppose we claim that the event that I am thinking about causation causes the event that I am typing this section. Now think about me thinking about causation, but not typing at the same time. Is that possible? My experience is that I did that a few minutes ago, so it is not necessary that I am typing this section if I think about causation. Therefore it is possible that it is not true that my thinking about causation causes me typing this. Yet one of the two possible causal claims that could be implied by the mentioned correlation is proved to be not necessary true.

The other claim is that the event that I am typing this section causes the event that I am thinking about causation. Is it possible that I am typing this, without thinking about what I type or anything else that has to do with causation? I think it is. Maybe I have written this all down before and I am just retyping it, thinking about going home and get some sleep. So the second possible causal claim is not necessary true too. In this case, there is correlation, but not necessarily causation. Therefore we can conclude that correlation does not imply causation, and thus that correlation is not the same as causation.

Correlation can *indicate* causation. If two events are correlated, it is possible that one of them causes the other. In our example, it is not necessary, but still possible that the fact that I am thinking about causation causes the fact that I am writing this (or the other way around). It is also possible that two correlated events cause each other, or that there is another event that causes them both at once. These last two kinds of causes will be discussed a bit further in Sections 2.3.3 and 5.2.

2.2 Categorizing theories of causality

We can categorize theories of causality according to the way they answer a number of key questions. This is Jon Williamson's approach in his chapter in *The Oxford Handbook of Causation* [Beebee et al., 2009, Chapter 9]. He divides these key questions into two categories, namely (a) questions that concern the causes and effects related by causality (i.e. the relata) and (b) questions that concern the causal relation itself. So if a causal relation is notated as $A \rightarrow B$, then (a)-questions are about A and B and (b)-questions

are about $\rightarrow$.

For the first kind of questions we can ask (a1) whether the causes and effects are single-case or generic, where with single-case we mean if the relata concern only a single situation (so they either obtain or fail to obtain) and with generic relata we mean that they can obtain and fail to obtain in different situations. We can also ask (a2) whether the relata at population-level or individual-level (is it about a group of individuals or just one individual?). The difference between single-case and generic relata is the difference between 'the smoking of my brother caused his cancer' and 'smoking causes cancer'.

The example that Williamson gives of an individual-level situation with single-case cause and effect is 'Audrey's letter will cause Balthasar anguish when he reads it'. The effect here is single-case, because it is the case that Balthasar gets anguish when he reads the Audrey's letter or it is the case that he will not. There are no two versions of Balthasar of which one gets anguish and the other does not, because with 'Balthasar' we mean a specific man ('Balthasar' is a uniquely identifying description). The cause 'Audrey's letter' is not a situation but an object, but I will interpret it as 'the reading of Audrey's letter by Balthasar', which concerns only a single situation, so this is single-case too. Even if Audrey sent more letters to Balthasar and Balthasar read them, we talk about a specific one of them here. The relata are also individual-level, because they are about one individual and not about a group of individuals.

As an example of a population-level with single-case cause and effect, the statement 'an increase in inequality of wealth in Britain in 2007 caused a reduction in happiness' is given. However, I disagree with Williamson that the effect here is single-case. Maybe the 'reduction in happiness' can be generic, because because this could obtain and this could fail to obtain at the same time for different people. Even for the generic example 'smoking causes cancer', there may be people that smoke without getting cancer. The same holds for ' an increase in inequality of wealth in Britain in 2007 caused a reduction in happiness', where there could have been people for which there was no reduction in happiness. I would say that the example Williamson gives about the reduction in happiness describes a causal relation between population-level a single-case cause and a generic effect.

For the second category (about the causal relation itself), we can ask (b1) whether the specific theory of causality says the connection between cause and effect is physical or purely mental (not physical). Besides that, we can ask (b2) whether the causal relation is objective or subjective and (b3) whether the theory attempts to understand actual causation (the factual

case) or potential causation (the general case).

About the questions of category (a), the questions concerning the relata, Williamson writes the following: "Of course *all* these kinds of causal relata occur in our causal claims, apparently without any great problem, so any theory that considers one kind to the exclusion of the other kinds provides only a partial account of causality." I agree with him that a theory that does not include all kinds of causal relata described in questions (a1) and (a2) is not a complete theory of causality. This means that from this categorization, one can infer whether a theory is complete or not.

2.3 What are problems when defining causation?

Many people have tried to give an overview of the problems that occur when defining causality. It is a fact that there is a lack of consensus in defining causation. We will see some reasons for this and look short at a fundamental problem of causal inference. Then we will consider problems there are with regularity theories of causation and compare this to the probabilistic approach, which turns out to be able to solve some of the problems of the first kind of theories.

2.3.1 Reasons for the lack of consensus in defining causation

In the Oxford Handbook of Causation, five reasons for the lack of consensus about the definition of causation are described [Beebee et al., 2009, Introduction, pp. 1-2]. The first is the fact that there are a great number of different theories of causation and the debate constantly continues, so it is hard to believe that an univocal analysis of the concept of causation is possible.

The second one is about the metaphysical status of causation. This is about the questions 'Is causation something fundamental in nature or not?' and 'Is it a feature of reality at all?'. (For further reading about the metaphysics of causation, see [Schaffer, 2008].) Even about that, there is no consensus.

The third reason is that philosophical theories of causation are derived from developments in the sciences in another way than most philosophical theories, because many philosophical theories of causation draw upon the resources of mathematical and scientific theories in their formulation. For example, probabilistic theories of causation draw upon the resources of probability theory in mathematics.

The fourth reason for the lack of consensus about causation is the fact that the concept of causality is used in many different contexts which do not

necessarily overlap. For example, causality is used in different contexts if we talk about the cause of the breaking of a window after throwing a stone in its direction, or talk about the statement that 'marriage causes less crime than being single' (where by 'being single' we mean 'not being married'). These are different types of causes, which we hardly can combine in one concept of cause.

The last mentioned reason is the central role of the concept of causality within many fields in philosophy, like epistemology, metaphysics and ethics. It turns out to be very hard to choose a definition of causation that applies to all areas.

2.3.2 A fundamental problem of causal inference

Something that might also be a reason for the lack of consensus about causality, is the following problem. It is impossible to measure two values of two different versions of a response variable for the same unit from the population, which is exactly the thing which describes the causal effect of the treatment on the unit. For example, it is impossible to look at the same man at the same time for two situations, namely the situation that he is married and the situation that he is not. You want to keep all other parameters, like age and total committed crime in the past, the same for the two situations. So it seems impossible to measure the effect of being married on men from the population, and therefore impossible to say something about the causal effect of being married. P.W. Holland called this the 'Fundamental Problem of Causal Inference' [Holland, 1986].

Let us discuss this in a more mathematical language, like Holland [1986] does it. We call the population of men U and one man from U is called u. The function $Y : U \to \mathbb{R} : u \mapsto x$ is a response variable (the part 'response' is because it is a variable that we want to explain), where $Y(u)$ for example gives us the amount of crime a unit u committed last year (we could think of a certain function which gives each kind of crime a mark of seriously and adds these). Now we make two different functions Y, namely Y_t to calculate the value of crime in the case that the unit is married and, Y_c, which stands for the case that the unit is not married (t stands for 'treatment' and c for 'control'). The effect of the cause t on unit u as measured by Y and relative to cause c is $Y_t(u) - Y_c(u)$ (causes of effects are always relative to other causes). What Holland now calls the fundamental problem of causal inference is the following. "It is impossible to observe the value of $Y_t(u)$ and $Y_c(u)$ on the same unit, and, therefore, it is impossible to observe the effect of t on u." [Holland, 1986, p. 947]

Two possible solutions for this are a scientific one and a statistical one. In a scientific experiment, it is not uncommon to assume that the time at which the experiment takes place has no influence on the result. In that case, it is possible to compare $Y_c(u)$ and $Y_t(u)$ when measured at different moments. So if a scientist is convinced of a certain value of $Y_c(u)$ (because he measured it several times), he can expose u to the treatment, measure $Y_t(u)$ and compare this with $Y_c(u)$.

A statistical solution to the problem is to calculate the average causal effect $T = E(Y_t - T_c)$, where all $u \in U$ are taken into account. It is known that $E(Y_t - Y_c) = E(Y_t) - E(Y_c)$ and $E(Y_t)$ and $E(Y_c)$ can both be calculated for respectively the u's that were exposed to t and the u's that were not. For these solutions, some typical assumptions are often used, which are mentioned in Section 4 of [Holland, 1986].

2.3.3 Problems of regularity theories

Let us first consider the kind of theories of causality that is called the regularity theories. This kind of theories is a legacy of David Hume, who wrote "we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second" [Hume, 2000, p. 46]. This approach does correspond to one side of the first intuition mentioned in Section 2.1.2, although we will see here the problem with the other side of that intuition (that there can be exceptions on a causal relation).

Hitchcock [2011] mentions some problems with this view on causation. First there is the problem of *imperfect regularities*. This says that most general causes have exceptions, so there are hardly any cases in this definition of cause. To look at our example, it is the case that marriage causes less crime than being single, but maybe not all people who marry really commit less crime than they would do when they would stay single. So not all events like C (someone who marries) are regularly followed by events like E (someone who commits less crime then before). Imperfect regularities present a problem for every theory of causality in which the definition of cause C has an element in the form 'for all x like C', because that does not allow any exceptions.

Another problem of the regularity view on causation is *irrelevance*. There are a lot of accidental correlations between events that are non-causal. For example, there could be an experiment which concludes that people who keep their oranges in the fridge, never have smelly feet. But to say that keeping your oranges in the fridge causes the lack of smelly feet, is probably

not right [Romeijn, 2010].

The third problem is that the causal relationship is *not symmetric*. With the previous definition of cause, there seems to be distinction between 'A causes B' and 'B causes A', because A can not be a cause of B if A is not followed by B. But there could be a causal circle in which A causes B and B causes A again, so in a theory that matches with this idea, such things must be possible in the definition of cause.

At least we have the problem of *spurious regularities*. Spurious regularities are relations between events with a *common cause* (events A and B are both caused by C). If C is the cause of A and B, it seems to be the case that besides that, A causes B (and B causes A, A causes C and B causes C), because they all satisfy Hume's definition. But we do not want them all to be causal relations, because only two of them are.

So we have to change the definition if we want it to cover all causes and if we do not want to allow any non-causes to satisfy the definition.

2.3.4 Probabilistic view

Probabilistic theories do solve two of the problems that come up in the regularity theories. Let us first look at something about the properties that are specific for a probabilistic view on causation. As mentioned before, the general structure of a causal claim is 'C causes E' and C (cause) and E (effect) are relata. In a probabilistic theory of causation, these relata are represented by events in a probability space. In this way, cause and effect become probabilistic dependent. An event is a cause if it changes the probability of another event (the effect). In (equivalent) formula's: $P(E \mid C) > P(E)$ or $P(E \mid C) > P(E \mid \neg C)$ [Hitchcock, 2011].

This solves the problem of imperfect regularities from the regularity theories. The exceptions of every cause just make a little difference in the probabilistic dependencies, but it remains a fact that there is dependence between the relata. The problem of irrelevance is also solved, because it is required that a cause makes a difference for the probability of its effect. But the problems of no symmetry and spurious correlations still remain. For the last problem, if C is a common cause of A and B, then C is called the confounding factor.

Sometimes, we want to take into account the moment in time at which the events take place. If event C happens at time t we write C_t. Hans Reichenbach gives a definition of cause that solves the problem of spurious correlations [Reichenbach, 1956]. He first introduces a definition of 'screening off'.

Definition 1. *C **screens** A **off** from E if and only if* $P(E \mid A\&C) = P(E \mid C)$.

This is the case in two different causal structures, namely if A causes C while C causes E and if C is a common cause of A and E. Then Reichenbachs definition of cause is the following:

Definition 2. *For events* C_t *and* $E_{t'}$ *with* $t' > t$, C_t *is a **cause** of* $E_{t'}$ *if and only if*

1. $P(E \mid A \cap C) = P(E \mid C)$ *and*
2. *there is no event* $B_{t''}$ *with* $t'' \geq t$ *such that* $B_{t''}$ *screens* $E_{t'}$ *off from* C_t .

Because of the second item, there only exist causal relations without a common cause that could possibly make the relation non-causal.

This notion of 'screening off' is important, because it is found in a lot of other books and articles. The first definition in Pearls book is that of Conditional Independence Pearl [2009]. X and Y are said to be conditionally independent given Z if $P(x \mid y,z) = P(x \mid z)$ whenever $P(y,z) > 0$. This means the same as saying Z screens off X from Y. The notation Pearl uses comes from Dawid (1979) and looks like this: $(X \perp\!\!\!\perp Y \mid Z)$. $X \perp\!\!\!\perp Y$ means that X and Y are independent and the total means that X and Y independent conditional on Z.

2.4 Conclusion

In this section we briefly discussed causality, but a sound definition is still lacking. The reasons for the lack of consensus and some other problems remain.

Characteristics of causal relations were mentioned, as such they contain relata, namely a 'cause' and an 'effect', which have to be clearly distinguished. Besides that, it is rendered plausible that correlation is not the same as causation nor does correlation imply causation. There was also described a way in which one could categorize theories of causality, namely by the way the theories answer key questions about the relata and about the relation itself.

Section 2.3 provided some reasons for the lack of consensus in giving a definition of causation and a fundamental problem of causal inference was mentioned. Problems with regularity theories were also discussed and in a short description of the probabilistic view on causality some of them appeared to be solvable.

In the next section, we will look at probability theories of causation in more detail.

3 Causal relations in terms of probabilities

We have seen the idea of probabilistic theories of causality. To use the probabilistic theory of Pearl [2009], we need to understand some terms involving graphs. First, some things about graphs are explained, because some kind of graphs is used a lot in modeling causal relations. Then there is some information about forks, d-separation and a do-operator we need to know, before we will see how we can go from the graphs to probabilities of relata in causal relations.

3.1 Directed acyclic graphs

Graphs are useful to provide convenient means of expressing substantive assumptions. Beside that they give an economical representation of joint probability functions and an efficient way to draw conclusions from observations. [Pearl, 2009, Section 1.2.2]

A graph G consist of a set vertices $V = \{v_1, \ldots, v_n\}$ and a set edges $E = \{(v_i, v_j), \ldots\}$. Given the finite set of vertices, the set of edges is also finite. While modeling causal relations, we use arrows to indicate which item influences the other. In that case, the edge (v_i, v_j) is directed towards v_j, so v_i is influencing v_j. See Figure 1 for the example in which $G = (V, E) = (\{v_1, v_2, v_3, v_4\}, \{(v_1, v_2), (v_1, v_4), (v_2, v_3), (v_3, v_1), (v_3, v_4)\})$.

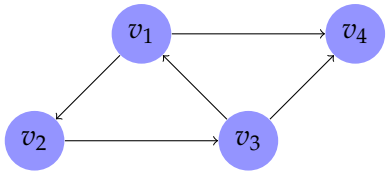

Figure 1: Example of a graph

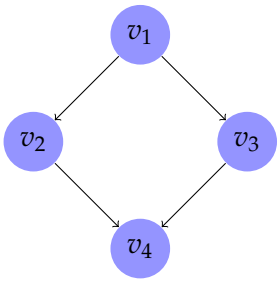

Figure 2: A directed acyclic graph

The kind of graphs used a lot in causal modeling are directed acyclic graphs (DAGs), see Figure 2 for an example. A graph is called directed when all edges of the graph are directed. That is the case when all edges are marked by an arrow. A graph is called acyclic if it does not contain any directed cycles. A directed cycle is a path (a sequence of edges such that each edge starts with the vertex ending the preceding edge) that is directed

(with arrows pointing from the first to the second vertex of each edge) and starts and ends with the same vertex. An example of a directed cycle path in Figure 1 is $((v_1, v_2), (v_2, v_3), (v_3, v_1))$ or $v_1 \to v_2 \to v_3$. Unlike Figure 1, which contains a directed cycle path, Figure 2 appears to be DAG.

Some notation needs to be explained because it will be used in this text. If (v_i, v_j) is a directed edge, then v_i is called the *parent* of the *child* v_j. *Ancestors* of a vertex are all vertices which are directly or indirectly linked to it with (an) arrow(s) pointing towards that vertex, so the ancestors of v_4 are v_1, v_2, v_3 in Figure 3. The *descendants* of a vertex are all vertices which are directly or indirectly linked to it with (an) arrow(s) pointing from that vertex away, so the descendants of v_1 are v_2, v_3, v_4. A graph is called a *connected graph* if there is a path between every pair of vertices.

3.1.1 Causal structures

A DAG can be used to represent causal structure. Pearl's definition of causal structure is as follows:

Definition 3. *[Pearl, 2009, p. 44] A causal structure of a set of variables V is a directed acyclic graph in which each vertex corresponds to a distinct element of V, and each link represents a direct functional relationship among the corresponding variables.*

Directed graphs are also known as Bayesian networks [Pearl, 1985]. Figure 2 can be a representation for the follow causal structure. Let us assume actual causation for marriage reducing crime. We can say v_1 stands for *marriage*, so v_1 is a binary variable ($v_1 = 1$ for someone that is married and if not, $v_1 = 0$), and v_4 stands for *crime* which we consider as a binary variable too. Two possible mechanisms for the influence of marriage on criminal behavior could be via *social commitment* and *everyday routine* [Sampson et al., 2006]. Let us measure these to variables with a mark on a scale from 1 to 5, so for example someone that has no social commitment at all, gets a 1 for that variable, and if he has a very regularly everyday routine, he gets a 5 for that one. Although it could be the case that the effect of marriage is not directly causal, we will use this example to explain causal structures. The causal structure is shown in Figure 3. It represents the assumption that the state of marriage of a person infleunces his social commitment and his everyday routine and it shows the assumption that these last two influence the amount of crime that the person commits (which results in a 0 if there is no crime and a 1 if there is).

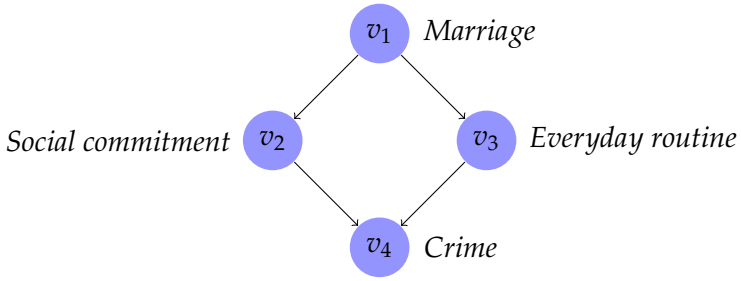

Figure 3: Example of a causal structure

3.1.2 Forks

D-separation is an important concept in causal modeling. To define this, we need to know something about forks and chains.

There are two different kind of forks in graphs that need to be discussed, namely the fork and the inverted fork. A *fork* is a set of three vertices of which one is linked to the other two by directed edges towards the other vertices. For example, $v_2 \leftarrow v_1 \rightarrow v_3$ in Figure 2 is a fork. The other kind of fork is a *inverted fork*, also called collider. This is a set of three vertices of which one is linked to the other two by directed edges that are directed toward the central one. An example of an inverted fork is $v_2 \rightarrow v_4 \leftarrow v_3$ in Figure 2.

A *chain* is a connected DAG in which every vertex had at most one parent and at most one child. Figure 2 would be a chain if v_2 or v_3 (including the arrows linked to it) would be omitted.

3.1.3 D-separation

Using these forks and chains, we can define d-separation. D-separation is a way to assign independent conditionality to sets of variables and is defined in the follow way:

Definition 4. *[Pearl, 2009, p. 16] A path p is said to be **d-separated** by a set Z of vertices if and only if*

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

2. *p contains an inverted fork $i \rightarrow m \leftarrow j$ such that the middle vertex m is not in Z and such that no descendant of m is in Z.*

A set Z of vertices is said to d-separate X from Y if and only if Z d-separates every path from a vertex in X to a vertex in Y.

For example, look at $X = \{v_1\}$ and $Y = \{v_4\}$ in Figure 2. These are d-separated by $Z = \{v_2, v_3\}$, because both paths between v_1 and v_4 are d-separated by a vertex in Z. The path $v_1 \to v_2 \to v_4$ is d-separated by Z, because it is (so contains) a chain $i \to m \to j$ such that the middle vertex m (in this case v_2) is in Z. The same holds for the path $v_1 \to v_3 \to v_4$, except that the middle vertex m is $v_3 \in Z$. The path $v_1 \to v_3 \to v_4$ is not d-separated by $Z = \{v_2\}$, because the middle vertex m would not be in Z. Because in that case there would be a path from v_1 to v_4 that was not d-separated by Z, X and Y would not be d-separated by Z.

If Z d-separates X from Y, than Z d-separates Y from X, because the set of 'paths from a vertex in X to a vertex in Y' is the same as the set of 'paths from a vertex in Y to a vertex in X'.

For another example, which uses the second item of Definition 4, we take $X = \{v_2\}, Y = \{v_3\}, Z = \{v_1\}$. Here there are again two paths from a vertex in X to a vertex in Y, namely $v_2 \leftarrow v_1 \to v_3$ and $v_2 \to v_4 \leftarrow v_3$. The first path is d-separated by v_1, because it is a fork such that the middle vertex is $v_1 \in Z$. The second path is a inverted fork, the middle vertex v_4 is not in Z, and v_4 has no descendants that could be in Z. Because of the second item of Definition 4, this path is also d-separated. Both paths are d-separated, so $X = v_2$ and $Y = v_3$ are d-separated by $Z = v_1$.

This d-separation "permits us to determine by inspection which sets of variables are considered independent of each other given a third set, thus making any DAG an unambiguous representation of dependency" [Pearl, 1988, p. 116].

3.1.4 Do-operator

In a causal structure, one can do an intervention that makes one of the relations (arrows in the graph) disappear. Suppose we have our group of criminal men and as intervention we make them all having a regularly everyday routine. We measured that with a mark on a scale from 1 to 5, so doing the intervention, we make all men get a 5 for 'everyday routine' (we suppose there is a way to do that, although I can not think of one). More about this is explained in the Section 3.2.

3.2 From graphs to probabilities

In Section 3.1.1 causal structures were defined. Here, we will define causal Bayesian networks, in which the do-operator is adjusted to tell something about the changing of probabilities. For the causal structure we already had, we can make a decomposition of probabilities in the following way. The probability of all variables ($P(v_1, ..., v_n)$) is the product of the probabilities for each vertex, dependent on their parents ($\prod_{1 \leq i \leq n} P(v_i \mid pa_i)$, where pa_i are all parents of v_i). So for Figure 3 the decomposition of $P(v_1, v_2, v_3, v_4)$ is $\prod_{1 \leq i \leq 4} = P(v_1)P(v_2 \mid v_1)P(v_3 \mid v_1)P(v_4 \mid v_2, v_3)$.

If we now do the intervention in which we make the value of v_3 equal to 5 for every men, we effectively make $P(v_3)$ independent of everything and $P(v_3 = 5)$ will be equal to 1. The new decomposition becomes $P_{v_3=5}(v_1, v_2, v_4) = P(v_1)P(v_2 \mid v_1)P(v_4 \mid v_2, v_3 = 5)$. The arrow between v_1 and v_3 can be removed, because v_3 no longer depends on anything. The new graph is shown in Figure 4.

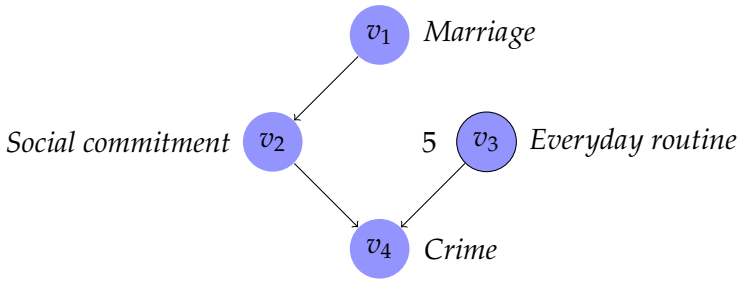

Figure 4: Example of an intervention in a causal structure

With this example, we go to the definition of causal Bayesian networks. The notation in Pearl's definition includes some different letters than we used above, but we will see them explained after the definition is given.

Definition 5. *[Pearl, 2009, p. 23-24] Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to constants x. Denote by P_* the set of all interventional distributions $P_x(v)$, $X \subseteq V$, including $P(v)$, which represents no intervention (i.e., $X = \varnothing$). A DAG G is said to be a **causal Bayesian network** compatible with P_* if and only if the following three conditions hold for every $P_x \in P_*$:*

1. *$P_x(v)$ is Markov relative to G;*

2. $P_x(v_i) = 1$ *for all* $V_i \in X$ *whenever* v_i *is consistent with* $X = x$*;*

3. $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$*, i.e., each* $P(v_i \mid pa_i$ *remains invariant to interventions not involving* V_i

In our example, the set of variables and the first probability distribution are $V = \{v_1, v_2, v_3, v_4\}$ and $P(v) = P(v_1, v_2, v_3, v_4)$. We only had one intervention, $do(v_3 = 5)$, which sets subset $X = v_3$ to constant $x = 5$, so $P_X(v)$ from the definition is $P_{v_3}(v_1, v_2, v_3 v_4) = P_{v_3}(v_1, v_2, v_4)$. Then P_* contains two elements, namely the distribution for $do(v_3 = 5)$ (which is $P_{v_3}(v_1, v_2, v_3 v_4)$) and the one without any intervention ($P(v)$). So we have $P_* = \{P_{v_3}(v_1, v_2, v_3, v_4), P(v)\} = \{P(v_1)P(v_2 \mid v_1)P(v_4 \mid v_2, v_3 = 5), P(v_1)P(v_2 \mid v_1)P(v_3 \mid v_1)P(v_4 \mid v_2, v_3)\}$.

Now, the graph $G = (\{v_1, v_2, v_3, v_4\}, \{(v_1, v_2), (v_1, v_3), (v_2, v_4), (v_3, v_4)\})$ is a causal Bayesian network compatible with P_* if and only if the three condition in Definition 5 hold for $P_{v_3}(v_1, v_2, v_4)$. The first condition says that $P_{v_3}(v_1, v_2, v_4)$ has to be Markov relative to G. This means that $P_{v_3}(v_1, v_2, v_3, v_4$ has to admit the factorization $P(v_1, ..., v_n) = \prod_i P(v_i \mid pa_i)$ relative to G. This is the case, because we used that formula to compute $P_{v_3}(v_1, v_2, v_3, v_4) = P(v_1)P(v_2 \mid v_1)P(v_3 = 5)P(v_4 \mid v_2, v_3 = 5)$. The second condition says that $P(v_3) = 1$ of $v_3 = 5$, which is indeed the case. The last condition, which says that $P(v_3 \mid pa_3) = P(v_3 \mid v_1)$ does not change for interventions that involve v_3, is also satisfied, because if we had made another intervention that not involved v_3, there would not be a difference in pa_3 and thus none in $P(v_3 \mid pa_3)$.

Important is the difference between observing that $v_3 = 5$ and making $v_3 = 5$ by yourself. When observing it, we may wish to infer that there is marriage ($v_1 = 1$) and there is a lot social commitment ($v_2 = 5$) (because we only expect a regularly everyday routine for men that are married and we expect a lot social commitment for men that are married), while in the case one makes $v_3 = 5$ by himself, no such inferences should be drawn.

3.3 Conclusion

We have seen the definition of a directed acyclic graph (DAG) and looked at causal structures in which forks were used to explain d-separation, which made a DAG an unambiguous representation of dependency. Besides that, the example of marriage that causes no crime was worked out in a graph representation and the do-operator was used to define causal Bayesian networks.

4 The counterfactual view on causality

As shortly mentioned before, causality can be viewed in a counterfactual way. In Section 2.3.3 there was given a definition of cause from Hume, but he gives another definition right after the first one [Hume, 2000] which says that C is a cause of E if and only if 'if C had not been, E never had existed'. This definition is counterfactual, because it is of the form 'If C had not occurred, than E would not have occurred'. Counterfactuals are sentences that state something that is not the case, but could have been the case (or something that is the case, but could have been not the case) if something else had been different. If we use our running example, C could be 'John married' and E could be 'John commits no crime'. This states that if John had not married, then John would not be committing no crime (that is John would be committing crime).

Probably the most well-known counterfactual theory of causation is from Lewis. This theory is about single-case causes, because generic causes in the form of 'C-events cause E-events' can be interpreted in many different ways, which would make it unnecessarily complicated [Lewis, 1973b, p. 558]. More about this theory is described further in this section, but before that, we will look at different kinds of implication used in logic.

4.1 Different kinds of implication

We will consider three different kinds of implication, namely material, strict and causal implication. To understand how strict and causal implication work, we first need to understand some things about modal logic. Modal logic is an extension of propositional logic, which is a formal system where symbols like A and B represent propositions and get a truth value 0 if that proposition is false and truth value 1 if it is true. How we can work with this, will be explained in Sections 4.1.2 en 4.1.3. After that, we will have a short look at the causal implication.

4.1.1 Material implication

In non-modal logic, there are some theories that are counterintuitive to many people. An example of these is a sentence with implication like

$$(A \rightarrow B) \vee (B \rightarrow A) \tag{1}$$

which says in fact that for two arbitrary sentences A and B at least one follows from the other [Kradavisvo, 2010]. So if for example A means 'It

is snowing in California now.' and B is 'I am writing this thesis now.', it is either true that the snowing in California follows from me writing this thesis or it is true that me writing this thesis follows from the snowing in California. The fact that this sentence is true is checked by the following truth table.

A	B	$(A \to B) \vee (B \to A)$
0	0	0 1 0 **1** 0 1 0
0	1	0 1 1 **1** 1 0 0
1	0	1 0 0 **1** 0 1 1
1	1	1 1 1 **1** 1 1 1

Table 1: Truth table for $(A \to B) \vee (B \to A)$

The implication used in Equation 1 is called the *material implication*. In modal logic we have *strict implication*, which does not mean that something *follows from* another thing, but means that something *follows necessarily from* the other thing. In that case $A \to B$ does not mean '*it is not the case* that B is false while A is true', as it is in non-modal logic, but $A \to B$ means '*it can not be the case* that B is false while A is true'. In modal logic, there is also something that means 'something follows possibly from another thing'. Instead for 'necessarily' and 'possibly', modal logic can also be used for other expressions like 'It is obligatory that X' and 'It is permitted that X', where the rules are a little different, but we will not use these here.

Using this strict implication, it is no longer possible to use truth tables, because there is no way to define a truth table for the 'necessary'-part of it [Garson, 2009]. If we want to check whether $(A \to B) \vee (B \to A)$ is still true for the strict implication, we will have to know something about the theory of possible worlds. Let us first introduce a new symbol to handle this new kind of implication: $\Box(A \to B)$ means that B follows *necessarily* from A. So we want to check whether $\Box(A \to B) \vee \Box(B \to A)$ is true.

4.1.2 Possible worlds

Possible-world-models were introduced by Kripke [1959] and are now called Kripke-models. According to Kradavisvo [2010, p. 66], a *Kripke-model* is a triple $M = \langle W, R, I \rangle$, where $\langle W, R \rangle$ is a *model structure* and I is the *interpretation function* of M.

To understand this definition, we look at a set $W = \{w_i \mid 1 \leq i < \infty\}$ of possible worlds. Between these worlds, there is a accessibility relation R that

tells which worlds are accessible from which other worlds. If a certain world w_j is accessible from world w_k, we notate this as $w_k R w_j$. The relation R is a set of pairs of worlds ($R \subseteq W \times W$), so for example if $R = \{\langle w_1, w_2\rangle, \langle w_2, w_4\rangle\}$ and $W = \{w_1, w_2, w_3, w_4\}$, then is w_2 accessible from w_1, w_4 is accessible from w_2 and further there are no worlds accessible from other worlds from W. We call the pair $\langle W, R\rangle$ a *model structure* when $W \neq \varnothing$.

The *interpretation function* I is a function from $PL \times W$ to 0,1, where PL is the class of proposition letters. For example, if A is true in world w_1, then $I : \langle A, w_1\rangle \to 1$, but if A is false in that world, then $I : \langle A, w_1\rangle \to 0$.

4.1.3 Strict implication

We give now an example of a Kripke-model in which $\Box(A \to B) \vee \Box(B \to A)$ is not true. Suppose $M = \langle W, R, I\rangle$ where $W = \{w_1, w_2, w_3\}$, $R = \{\langle w_1, w_2\rangle, \langle w_1, w_3\rangle\}$ and $I(\langle A, w_1\rangle) = I(\langle B, w_1\rangle) = I(\langle A, w_2\rangle) = I(\langle B, w_3\rangle) = 1$ while $I(\langle B, w_2\rangle) = I(\langle A, w_3\rangle) = 0$. There can be drawn a graph of this, which is shown in Figure 5. Because $I(\langle A, w_2\rangle) = 1$ and $I(\langle B, w_2\rangle) = 0$, we know that $A \to B$ is false in w_2. Because of that and the fact that $w_1 R w_2$, we can conclude that $\Box(A \to B)$ is false in world w_1. In the same way, we know that $B \to A$ is false in w_3 (because $I(\langle B, w_3\rangle) = 1$ and $I(\langle A, w_3\rangle) = 0$) and $w_1 R w_3$, so we can conclude that $\Box(B \to A)$ is false in world w_1. Because both $\Box(A \to B)$ and $\Box(B \to A)$ are false in w_1, we have a possible world in a model in which $\Box(A \to B) \vee \Box(B \to A)$ is false. So in modal logic, where strict implication is used instead of material implication, the idea that for two arbitrary sentences A and B at least one follows from the other is not true at all possible worlds.

According to Lewis, a necessity "acts like a restricted universal quantifier over possible worlds" [Lewis, 1973a, p. 4]. Let us call a world in which A is true an A-world. Then $\Box(A \to B)$ is true in $w_i \in W$ if and only if for all $w_j \in W$ for which $w_i R w_j$ and A is true at w_j, then B is true at w_j. In other words, $\Box(A \to B)$ is true in $w_i \in W$ if and only if B is true in all accessible A-worlds [Lewis, 1973a, p. 5]. With this information, we can see clearly that $\Box(A \to B)$ and $\Box(B \to A)$ both are not true in w_1 in Figure 5 and if we know that, we know that $\Box(A \to B) \vee \Box(B \to A)$ is not true.

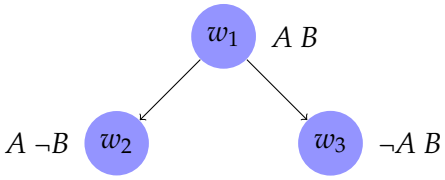

Figure 5: Example of a Kripke-model in which $\Box(A \rightarrow B) \vee \Box(B \rightarrow A)$ is false at world w_1

4.1.4 System of spheres

Corresponding to a necessity operator, there is an assignment to each world w_i of a set S_{w_i} of worlds, which is called the *sphere of accessibility* [Lewis, 1973a, p. 7]. We can say a sentence $\Box A$ is true at a world w_i if and only if A is true throughout the sphere of accessibility S_{w_i} around w_i. Using this new idea, $\Box A \rightarrow B$ is true at w_i if and only if for all worlds $w_j \in S_{w_i}$ in which A is true, B is also true. This means the same as saying that $\Box(A \rightarrow B)$ is true in w_i if and only if B is true in all accessible A-worlds.

We can make a set $\$_{w_i}$ of sets of possible worlds (Lewis uses a dollar sign with two vertical lines instead of one) and define a system of spheres in the following way:

Definition 6. *$\$ is called a **system of spheres**, and members of each $\$_{w_i} are called **spheres around w_i**, if and only if, for each world $w_i \in W$, the following conditions hold.*

1. *$\$_{w_i} is centered on w_i, which means $\{w_i\} \in \$_{w_i}$.*
2. *$\$_{w_i} is nested, which means if $S, T \in \$_{w_i}$, then $S \subseteq T$ or $T \subseteq S$.*
3. *$\$_{w_i} is closed under unions, which means if $\mathcal{S} \subseteq \$_{w_i}$ and $\bigcup \mathcal{S} = \{w_j \mid \exists X$ such that $w_j \in X \in \mathcal{S}\}$, then $\bigcup \mathcal{S} \in \$_{w_i}$.*
4. *$\$_{w_i} is closed under nonempty intersections, which means if $\varnothing \neq \mathcal{S} \subseteq \$_{w_i}$ and $\bigcap \mathcal{S} = \{w_j \mid w_j \in X \; \forall X \in \mathcal{S}\}$, then $\bigcap \mathcal{S} \in \$_{w_i}$.*

The things we have seen here can be used to define causal implication, as we will see in the next subsection.

4.1.5 Causal implication

Lewis uses the following symbols that can indicate a counterfactual: $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$. $\phi \Box\!\!\rightarrow \psi$ means 'if it were the case that ϕ, then it would be the case that ψ'. $\phi \Diamond\!\!\rightarrow \psi$ means 'if it were the case that ϕ, then it might be the case that ψ' [Lewis, 1973a]. Here, we will only use the first one. For example, the operation $\Box\!\!\rightarrow$ is defined as follows:

Definition 7. *[Lewis, 1973a, p. 16] $A \Box\!\!\rightarrow B$ is true (at a world w_i) if and only if one of the following items holds:*

1. *No A-world belongs to any sphere $S \in \$_{w_i}$.*
2. *Some sphere $S \in \$_{w_i}$ does contain a least one A-world, and $A \rightarrow B$ holds at every world in S.*

With this definition, a lot of work with causal relations can be done (some of which is described in [Lewis, 1973b]), but we will have to understand it much better before we can deal with that.

4.2 Causal effects in counterfactual logic

Counterfactuals can be true or false. The approach which Sampson et al. [2006] use to check whether marriage reduces crime, is a counterfactual method that conceptualizes causality in terms of the effect of the treatment 'marriage' on the outcome 'likelihood of committing a crime'. The sample population is divided in a treatment group (those who marry) and a control group (those who do not marry). It is assumed that theoretically each unit has two potential outcomes that are demonstrated under the treatment respectively the control condition: Y_i^t and Y_i^c. These two outcomes cannot be observed at the same time for the same individual. That is a problem that is solved in experimentation through randomization. "Assuming equivalence of controls and treatments (...) permits the estimation of the causal effect, $Y_t - Y_c$" [Sampson et al., 2006, p. 472]. The method of 'propensity score matching' is used. With this method, the propensity of getting married is modeled for each individual and the population is divided into a group with individuals that get married on this propensity score and a group with individuals that do not get married on this propensity score. More about this method can be read in Morgan and Winship [1999]. Sampson et al. also used the method of inverse probability-of-treatment weighting (IPTW) (which they mention on page 473 and further).

4.3 Conclusion

In this section, we have looked at different kinds of implication and we showed that with material implication some sentences that we regard as nonsense can be true. The example we gave was the sentence $(A \to B) \vee (B \to A)$, which states that for two arbitrary sentences A and B at least one follows from the other. It was showed that using strict implication, it is possible that this sentence is false.

Besides that, we have seen how Lewis defined causal implication with counterfactuals, although we did not work that out very extensive, and in the end, we looked very short at the counterfactual method Sampson et al. used to draw causal conclusions in their study.

5 The problem of confounding

We have seen many problems with theories about causality in Section 2, but here will be looked at confounding, which is about factors that influence the results of a study while the researchers do not want them to have that effect or do not know these factors exist. We look at the expectations a researcher has from his results when there would be a causal relation between the investigated variables and things that can go wrong when someone concludes a certain causal relation from the viewed data.

5.1 What is confounding?

In Section 2.3.4 we have already seen that if C is a common cause of A and B, then C is called a confounding factor.

According to Greenland et al. [1999], there are at least three different definitions of confounding. I will mention two of them. In non-experimental research, confounding is used in the sense of being a type of bias in estimating causal effects. Informally, this is a combination of the effects of confounders and the effect of interest, whereby with confounders extraneous factors are meant.

In experimental-design literature, confounding is used as reference to inseparability of main effects and interactions under particular design. In literature about analysis of variance this is called aliasing.

One of the authors of Greenland et al. [1999] says something about confounding in his book about causality: "Confounding is a simple concept. If we undertake to estimate the effect of one variable (X) on another (Y) by examining the statistical association between the two, we ought to ensure that the association is not produced by factors other than the effect under study." [Pearl, 2009, p. 182]

5.2 Difficulties of causal inference

In his *Fundamentals of Statistical Causality*, Philip Dawid enumerates seven difficulties of causal inference, specifically in statistical problems. He says "all the difficulties go under the general head of confounding - meaning that the observed "signal" in the data need not be a pure effect of the putative cause under study, but may also be due, in whole or in part, to other factors that vary together with that "cause"." [Dawid, 2007, p. 14]. Here are the difficulties Dawid describes:

1. We have seen the first of the difficulties Dawid mentions in Section 2.3.3, namely the problem of common causes. In the research of marriage reducing crime [Sampson et al., 2006], the selection of units is nonrandom because the population on which the research was done existed of men that as adolescents had a high risk of being a criminal in later life. There could be a common cause for committing no more crime and getting married, for example moving to a new place, where there are no more old friends that made you committing crimes and bullied you when you wanted to date a girl.

2. There is *complete confounding* if together with the expected cause there is something else different in the cases with and without the effect. Dawid gives the example of someone who drives to his work using Shell petrol and back home using BP fuel and thinks that BP petrol gives better fuel economy than Shell petrol, while the way to work is uphill and the way back downhill. The uphill/downhill fact and the use of Shell petrol or BP fuel can both logically be causes of the difference in fuel economy. Therefore you can't just draw a causal conclusion from this situation.

   It is difficult to say whether there is complete confounding in the study of Sampson et al., because it is hard to determine if there is something else than marriage that is a potential cause of committing no more crime and that is different in the cases that there is no crime anymore and the cases that there is still crime. Maybe everyone that is married has been in the council-house and perhaps that has a huge influence on their behavior.

3. *Reverse causality* is the problem that in a claimed (observed) causal relation between two relata, it is possible that the claimed effect is in fact the cause and the claimed cause is in fact the effect. There is *two-way causality* if the two relata affect each other. In that case, both the relata are cause and effect. In Section 5.3 you can see an example which might be reversed causality. In our example of marriage causing no crime, there could be reverse causality, if the men that stop committing any crime are more likely to get married (maybe because they start being more open against the people in their neighborhood).

4. When choosing a population to examine, the problem with *selection* is that the units from the chosen population are not representative. They could have special common properties that cause the measured effect, while the researcher thinks the effect is caused by the treatment.

5. Another example Dawid gives is the following. Three years ago there were 62 fatal accidents (in total) at some specific locations and after placing some red light cameras the number of accidents this year is 19. If this fact makes us decide to place red light cameras for all traffic lights in order to a decrease of accidents, we forget considering *regression to the mean*. Even if there were no red light cameras installed, we would still expect a decrease of accidents (because the measurements before placing the cameras were the result of random fluctuations). We should first compare the new measurements from the locations with cameras with measurements from locations without cameras. If the latter give the same result as the former, it is unlikely that the number of accidents decreased because of the red light cameras.

   There is no problem of regression to the mean in the criminology-example, because where in the example of that there were only measured sites with cameras and none without, here the men that get married were taking into account just as much as the man that did not marry.

6. We can view data of the efficiency of medicine by comparing the group with treatment with the group without treatment. But we could also split the population in for example men and women, and compare the ones with and without treatment in both groups apart. If the first comparison shows that treatment increases the chance on recovering, it is still possible that the second comparisons show that for men and for women the chance on recovering decreases with treatment. This is known as *Simpson's paradox*. For a detailed example, see [Dawid, 2007, p. 13]. This paradox does not apply to the study after marriage causing no crime.

7. To explain the idea of the problem with *promotion and prevention*, we look at the following example from Dawid. When someone compares students with a fake ID with the drinking habits of students, and if he thinks about interfering with the experiment by taking and giving away fake IDs, he has to realize the difference in the amount of effect these two actions have. It is expected that taking someone's fake ID has more effect on someone's drinking habits than giving one to someone without a fake ID. This has not been a problem for the criminology-example, but it could have been if the study was done in another way. Suppose that in the experiment young men from prison that have no girlfriend yet are coupled to some young women that

are available for marriage and these couples do indeed marry. On the other hand, there are young men that are already married and in some way, we make them get a divorce. I think there would indeed be another reaction of the ones that were given marriage than the ones of which their marriage was taken away.

5.3 Statistical example of confounding

Chang et al. [2011] have tried to show that in the elderly in Taiwan frequent shopping increases survival. My intuition says this has to be the other way around, namely that survival in the elderly causes more frequent shopping, or that there is a common cause like 'being a healthy person' that causes both 'often shopping in later life' and 'becoming very old'. Chang et al. admit that the relationships of social activities to survival are confounded by functional health. To prevent the influence of this confounding, they made an extra model that used a sub-sample where participants with great difficulty in shopping were excluded. We will discuss some details of this study and look at a statistical model for analyzing data that was used, namely the Cox proportional hazard model.

5.3.1 Details of the study

Shopping frequency, cognitive function, physical function and many other items were measured for a group of 1841 Taiwanese people with a minimum age of 65. On the question 'What is the frequency with which you go out shopping?' were four possible answers: 'never or less than once a week', 'once a week', 'two to four times a week' and 'every day'. I wonder whether Chang et al. looked only at weekdays or also at days in the weekend. I believe shopping in Taiwan is possible seven days a week, so if they looked at all days of the week, I wonder why they did not state five instead of four in 'two to four times a week' and 'six to seven times a week' instead of 'every day'. Literally, the difference shopping 'two to four times a week' could be done in one day, while someone that answered 'every day' could be shopping continuously. The formulation used by the researchers does suggest that they did not take into account whether or not people did go shopping more times a day, so someone that went shopping three times every Saturday (so he can stay inside for the rest of the week for example) would be in the group that shopped once a week. If shopping really increases the age of dying, shopping three times on one day a week would probably have a bigger influence on your age of dying than shopping just

one time a week. Even if such a person was in the group of people that shopped 'two to four times a week' he might have had another chance of dying than someone that shops one time a day on three days a week. I could give more of such examples of why I think this choice of possible answers is not ideal, but assuming these special situations are rare exceptions I will assume this is not influencing the results of the study.

Chang et al. are looking at the influence of shopping freuqency on survival, but if they conclude that the more a elder person shops, the longer he or she will live, it is not necessary that this is because of the shopping itself. It can be that with shopping comes something else, like 'taking a walk', that is the actual cause of the increase of the survival time. They did not measure the money that was spend per shopping time, so even if you really would like it, you should not conclude that the more money you spend the longer you live.

What would Chang et al. expect as results if there was a causal relation between shopping and age of dying? They used Cox proportional hazard models to look at the connection between shopping frequency and survival time, where they used the group that shopped never or less than once a week as reference group. Survival time was determined as the time between the interview (in 1999 or 2000) and the date of death or 31 December 2008 for subjects that survived until then and later. In these Cox proportional hazard models potential confounders can be included as covariates, like gender, ethnicity and physical functioning, so that these are excluded from having influence on the potential causal relation. From these models hazard ratios can be obtained, which indicate how much the variable of interest, in this case shopping frequency, has an effect on the survival time of the subjects. If the hazard ratio is 1, there is no correlation at all between shopping frequency and survival time. The more different the hazard ratio is from 1, the more correlation there is and the more likely it is that there is a causal relation between the two things.

5.3.2 Cox proportional hazard model

Suppose you want to know the chance of dying for a certain person in a certain infinitesimal time interval. This can be calculated with the hazard function for Cox proportional hazard model:

$$\lambda(t \mid X) = \lim_{dt \to 0} \frac{P(event \in (t,\ t + dt) \mid no\ event\ untill\ t)}{dt} \tag{2}$$

In this formula, t stands for the time that has passed since the beginning of the research, X is the covariate vector and in this case 'event' means 'dying'. It gives the limit of the chance of dying within a time interval after t given that the person in question has not died before time t. To understand what the covariate vector is, we view the example in which we adjust a model for gender, age at baseline and education. If the person for who you want to know the chance of dying is a male of age 72 which has been in college, you have to look in Table 2 (which is a short version of Table 1 in [Chang et al., 2011]) to see in what kind of group the researchers divided the units, so you can determine the values in the covariate vector. The vector is of the form $X = ([x_1], [x_2], [x_3])$ where every x_i exists of as many zeros and ones as the number of different groups minus one, so for example x_3 stands for education and exists of three times a zero or one, because there are four groups of education in this research. Each zero or one stands for one of the groups and only the first group does not have a zero or one that refers to it. The group in which the considered unit is placed becomes a one and the other groups become zeros, so in each $[x_i]$ there is at most one one.

Gender
Male
Female
Age at baseline
65-69 y
70-74 y
75-97 y
Education
Illiterate
Elementary school and below
High school
College and above

Table 2: Some measured variables in [Chang et al., 2011]

The hazard function gives the hazard (the risk of dying) at time t for one unit. Because the person we look at has been in college, we make the last digit of x_3 a one. The rest of the digits in x_3 are zeros, so $x_3 = [0, 0, 1]$. If we do the same for gender and age, we get our covariate vector $X = ([0], [1, 0], [0, 0, 1])$.

According to Cox [1972], the following model can be used as hazard

function:

$$\lambda(t \mid X) = \lim_{dt \to 0} \frac{P(event \in (t,\ t+dt) \mid no\ event\ untill\ t)}{dt} = \lambda_0(t)\exp(X\beta) \quad (3)$$

Here X is the same covariate vector as before and in our example β is the columnvector $(\beta_{gender}, \beta_{age\ at\ baseline}, \beta_{education})$, where $\beta_{age\ at\ baseline} = (\beta_{70-74y}, \beta_{75-97y})$, $\beta_{gender} = (\beta_{female})$ and $\beta_{education} = (\beta_{elementary\ school\ and\ below}, \beta_{high\ school}, \beta_{college\ and\ above})$.

The hazard ratio, which can for example measure the effect of gender on the chance of dying, can be calculated by taking the ratio of the hazard function for a male and the hazard function for a female, where all other variables such as education are the same for the male and for the female:

$$HR = \frac{\lambda(t \mid X_{male})}{\lambda(t \mid X_{female})} = \frac{\lambda_0(t)\exp(X_{male}\beta)}{\lambda_0(t)\exp(X_{female}\beta)} = \exp((X_{male} - X_{female})\beta) \quad (4)$$

If we continue with our example, we have $X_{male} = ([0],[1,0],[0,0,1])$ and $X_{female} = ([1],[1,0],[0,0,1])$, so $X_{male} - X_{female} = (1,[0,0],[0,0,0])$. From this, we can conclude that the hazard ratio is the following:

$$HR = \exp((1,[0,0],[0,0,0])(\beta_{gender}, \beta_{age\ at\ baseline}, \beta_{education})) = \exp(\beta_{gender}) \quad (5)$$

This works the same if you take not (only) gender, age and education, but as many other variables as you want. The more variables you include, the more potential confounders you prevent from having effect on your potential causal relation.

Unlike what we just did, namely give a way to check whether gender would influence survival time, the Chang et al. wanted to know if shopping frequency had any influence on survival time. But this works the same, instead the names of the elements in β and the number of elements in X change. Chang et al. made some different models in which they used different covariates, some of which you can see in Tabel 3. This tabel shows the relationship between shopping frequency and risk of death for four models that included a different number of covariates. Model 1 is adjusted for gender, age, education, ethnicity, alcohol drinking, smoking, exercise, dinner companions, comorbidity and region. Model 2 is adjusted by model 1 covariates plus perceived financial status, work status and transportation. Model 3 in this table is model 5 in Table 3 of Chang et al., which is adjusted by model 2 covariates plus physical functioning and Short Portable Mental Status Questionnaire. At least, model 4 (model 6 from Chang et al.) is model 3 for a sub-sample, which excluded those who were unable to shop

due to difficulty. If we look at the graph I made from this table, we can see that all hazard ratios for 'never or less than once' are 1. This is because this group was used as reference group, so for this hazard ratio the denominator and the numerator are the same. It is remarkable that the hazard ratios for 'two to four times a week' are closer to 1 than the ones for 'once a week'. If we do not look at the ones for 'every day', this would mean that the more you shop, the less correlation there is with age of dying. I really think Chang et al. should have used other possible answers for the question after shopping frequency, because if they used for example an extra option 'five to six times a week', there probably should be a more convincing graph that showed a decreasing line from 'two to four times' via 'five to six times' to 'every day'. It would of course also be possible that there would be an even more non-convincing graph then the one we have here, if the hazard ratios for 'five to six times' would have been very close to one.

We see that model 1 to 4 look as if they are be reversed in order in the graph. By the description of these models, we have seen that from model 1 on, every next model is adjusted for more covariates than the last one, so the more covariates there are used, the closer the hazard ratios get to 1. A possible conclusion could be that the added covariates actually had some correlation with the age of dying. In 5.3.3 we will see why we not directly conclude it, but just call it a 'possible' conclusion.

The last model is the one that excluded the people that were very ill and unable to shop. Noteworthy is that in this case not only the hazard ratio of 'two to four times' shopping is close to 1, but also the one of 'once' is more away from 1 than the one for 'every day'. So I doubt if we can conclude that frequently shopping is correlated with age of dying.

		Weekly shopping frequency			
		HR (95% CI)			
	n	**≤ 1**	**1**	**2 – 4**	**Every day**
Model 1	1816	1	0.60	0.73	0.58
Model 2	1744	1	0.73	0.80	0.61
Model 5	1713	1	0.83	0.94	0.73
Model 6	1574	1	0.89	0.98	0.76

Table 3: The relationships between shopping frequency and risk of death, part of Table 3 in [Chang et al., 2011]

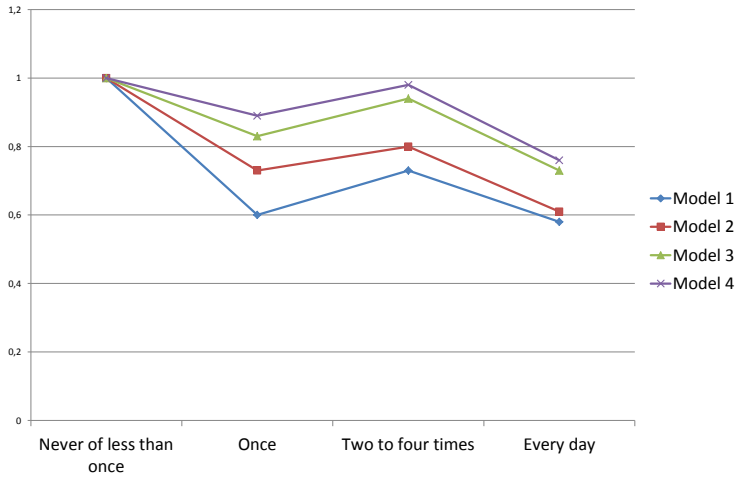

Figure 6: A graph of Table 3

5.3.3 Problems with Cox proportional hazard models

What are problems with this? Cox proportional hazard models assume that the hazard ratios are constant in time, so the HR needed only be computed at the end of the study. The HR could be different if it were computed at half of the time of the study, because less people would have died by then [Hernán, 2010]. Therefore it would be interesting to calculate hazard ratios at other moments of the study instead of only the moment at the end of the study. In that case the average HR could be computed, but that is not a solution at all, because if the hazard ratios for example 0.41 for the first and 1.56 for the last years of the study, the average hazard ratio will be close to 1. This would indicate that there is no correlation, while the hazard ratios of the two parts of the study tell there is (although it is in a bit weird way). We call the hazard ratios for the two parts of the study 'period-specific' hazard ratios. If we only look at these, we come to another problem with hazard ratios, namely the fact that they have a built-in selection bias. This means that something goes wrong in choosing the groups in the study that can not be prevented. For example, in a study after a certain disease where a part of the subjects gets treatment and a part of the subjects gets a placebo, the proportion of subjects that were susceptible to the disease at the start

of the experiment was unknown but expected to be the same in both the treatment and the placebo groups (because of randomization). However, with time, the proportion of susceptible subjects increases in the placebo compared with the treatment group. Why this is the case, can be read in [Hernán, 2010, p. 14].

5.4 Conclusion

We have seen some difficulties of confounding that Dawid gives in his report and for some of them we made our own examples. After that, we looked at the study of Chang et al. [2011] that tries to draw causal inferences of data via Cox proportional hazard models. My opinion is that Chang et al. conclude too fast that frequent shopping increases survival, because of the following:

- The statistics only say there is correlation, and if there would be causation, it could be that become old causes frequent shopping instead of the other way around.
- The model with the subsample that excluded participants with great difficulty in shopping is not convincing.
- It is unclear whether it was 'shopping', or something that comes together with shopping, that might be a cause of a longer live.
- The hazard ratios were computed only for the end of the study, so the result could have been very different is the study had been for a longer or shorter period of time.

6 Conclusion

Considering all theories and examples we have seen about causation, I can conclude several things.

First, there is a lack of consensus about causality and there are a lot problems with defining it. Therefore, we still do not understand exactly how causality works, although we can work with it if we accept certain theories. We can do a lot of things that have to do with causal relations, like drawing graphs and ad probabilities to it. Also logic that includes counterfactuals can be very interesting, although I still do not know exactly how a counterfactual theory of causation works.

Beside the lack of consensus about causation and the problems with defining it, there are also many problems in drawing causal inferences from statistical data. Some of them are specific for certain studies (for example the ones in which Cox proportional hazard ratios are used) and some of them are more general (for example the fundamental problem of causal inference that was mentioned in Section 2.3.2).

If I would have to choose which of the two major theories of causality I described is the best I would like to learn some more about them first. Especially about the counterfactual theories, I still have not a precise idea of how one can work with it. I understand more of the probabilistic theory than of the counterfactual one, but that does not make it a better theory. Nevertheless, if we measure the goodness of a theory according to how much it matches with our intuition, I could choose the probabilistic one because for the counterfactual one I do not know if it matches with my intuition. On the other hand, I do not know if it includes a lot of counterintuitive theorems, while I know the probabilistic theories still have some problems. I conclude that I will not choose before I have learned more about both kind of theories, especially about the counterfactual ones.

References

H. Beebee, C. Hitchcock, and P. Menzies. *The Oxford handbook of causation*. Oxford University Press, Oxford, 2009.

Y.H. Chang, R.C.Y. Chen, M.L. Wahlqvist, and M.S. Lee. Frequent shopping by men and women increases survival in the older taiwanese population. *Journal of Epidemiology and Community Health*, 2011. Published Online First: 6 April 2011 doi:10.1136/jech.2010.126698.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

A.P. Dawid. Fundamentals of statistical causality. Technical Report 279, Department of Statistical Science, University College London, 2007.

J. Garson. Modal logic. In *The Stanford Encyclopedia of Philosophy*. 2009.

S. Glueck and E. Glueck. *Unraveling Juvenile Delinquency*. Commonwealth Fund, New York, 1950.

S. Glueck and E. Glueck. *Delinquents and nondelinquents in perspective*. Harvard University Press, Cambridge, 1968.

S. Greenland, J.M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999.

M.A. Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010.

C. Hitchcock. Probabilistic causation. In *The Stanford Encyclopedia of Philosophy*. 2011.

P.W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(398):945–960, 1986.

M. Hulswit. *From cause to causation: A Peircean perspective*, volume 90. Kluwer Academic Publishers, Dordrecht, 2002.

D. Hume. An enquiry concerning human understanding: a critical edition. 3, 2000.

Y. Iwasaki and H.A. Simon. Causality in device behavior. *Artificial intelligence*, 29(1):3–32, 1986.

Phoebe Kradavisvo. Model, oneindigheid en paradox, syllabus en leesstukkenbundel, 2010.

S.A. Kripke. A completeness theorem in modal logic. *The journal of symbolic logic*, 24(1):1–14, 1959.

D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973a.

D. Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1973b.

S.L. Morgan and C. Winship. The estimation of causal effects from observational data. *Annual Review of Sociology*, 25:659–706, 1999.

J. Pearl. *Bayesian Networks: A Model of Self-Activated: Memory for Evidential Reasoning*. Computer Science Department, University of California, 1985.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, Inc., 1988.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, Cambridge, 2009.

H. Reichenbach. *The direction of time*. University of California Press, 1956.

J.W. Romeijn. Wetenschapsfilosofie minor college 4 "wetenschappelijke verklaringen", 2010.

R.J. Sampson, J.H. Laub, and C. Wimer. Does marriage reduce crime? a counterfactual approach to within-individual causal effects. *Criminology*, 44(3):465–508, 2006.

J. Schaffer. The metaphysics of causation. In *The Stanford Encyclopedia of Philosophy*. 2008.

A. Stevenson. *Oxford dictionary of English*. Oxford University Press, 2010.