Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Causal inference: An introduction

(Dutch title: Causale inferentie: Een introductie)

Thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

J. van der Ster

Delft, The Netherlands
July 2018

BSc thesis Applied Mathematics

"Causal inference: An introduction"

(Dutch title: "Causale inferentie: Een introductie")

J. van der Ster

Delft University of Technology

Supervisors

Prof.dr.ir. G. Jongbloed Dr. K.P. Hart

Thesis committee

Dr. B. van den Dries

July, 2018 Delft

Acknowledgements

First, I would like to thank my supervisors Geurt Jongbloed and Klaas Pieter Hart for their support and valuable guidance. Their expertise, knowledge and advice have helped me accomplish this paper.

I am indebted to my school colleagues who have helped make my learning an enjoyable and stimulating experience. I am very thankful to them for their encouragement.

Lastly I wish to thank my family and my close friends, whose enthusiasm, interest and support in this venture have given me the motivation to realize this achievement.

I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in this project.

Abstract

Experiments have always been the way to study what the effect is of interventions. Causal inference is an important aspect.

In this thesis we gave an introduction to causal inference. We did this by giving an example that illustrates the Fundamental Problem of Causal Inference. The Fundamental Problem of Causal Inference states that it is impossible to observe the values of control and treatment on the same unit and therefore impossible to observe the effect of the treatment on a unit.

We used a standard statistical model to later introduce the model for causal inference. The model we used for causal inference is Rubin's model. We assumed that there are two levels of treatment: control and treatment. Both are causes and we determine an effect of a cause always relative to another cause.

We discussed a range of assumptions to make it possible to estimate the causal effect. None of them are provable, the best we can do is convince ourselves and others of its correctness. We divided the solutions in two categories: the scientific solutions and statistical solutions.

The solutions were then used to investigate an issue about alcohol consumption in some newspapers. We concluded that there has to be more awareness about causal inference.

Contents

Introduction

The central question in many branches of science and technology is: "Does intervention have an effect?" For example, we want to investigate whether a drug has a positive effect on a patient. Or we want to know if fertilizer works on trees. Another example is whether a new way to learn in school works. We study these questions on the basis of data. These data are gathered using one of the two following methods: an experimental study or an observational study.

In an experimental setting, we set up an environment where we can manipulate a particular factor. We do this in order to provide insight into the effect of this factor. Usually, the experiments are in a laboratory setting to apply random assignment or to completely control confounding factors. Random assignment is the technique to assign people or other research objects to different groups by chance, for example by flipping a coin. Confounding factors are, simply put, factors that influence other relevant variables.

In an observational study however, we cannot manipulate a particular factor. This can be impossible, impractical or unethical. As such, we need to know and account for confounding variables. However, it is often difficult to control these variables, that is, to keep them constant.

An interesting compromise is the case-control study. This is in essence an observational study, but the main difference between the two is that a case-control study compares two existing groups with different outcome. The intervention of a scientist is thus after the exposure and after the disease.

The question that arises in all of these studies is: "Is there a causal effect?" Often times, we want to show that there is a causal effect or deny the fact that there is one. The government can use it to make scientifically justified decisions, but a toothpaste company can use it to show that their product is the best. But when can we conclude rightly that there is a causal effect?

Huff (1954) illustrated multiple ways to lie with statistics. Some are obvious, others are less apparent. One of the less obvious ones is wrongly drawing causal inferences. Holland, Glymour, and Granger (1985) were among the first to show that statistical models used to draw causal inferences are distinctly different from those used to merely describe a relationship between two data sets.

This thesis introduces causal inference to both mathematicians and to non-mathematicians, by delaying the mathematical part to a later section. It fills the gap[1] between the mathematical papers published about this subject and the needs of non-mathematicians or starting mathematicians.

In order to this, we will first introduce the problem of causal inference in a non-mathematical way. An example will provide a feeling for the Fundamental Problem, which will be discussed after the example. After that we will introduce mathematical methods and explain the difference between a model for associative analysis and a model for causal inference. In the end we will look at an example about alcohol consumption where several media interpreted a research report incorrectly.

[1]The concept of causal inference is hard to grasp for (non-)mathematicians and this thesis tries to explain the core of causal inference to them and in a later section explains the fundamentals to starting mathematicians.

1 New drug: Treatment vs Control

In medical research an aim is to increase the longevity of humans. One way to do this is to invent drugs that prevent or heal illness. After the inventing and developing of such a new drug, we want to know how this drug performs. Most people tend to forget that not every new drug is better than the last one. But how do you test which one performs better? The standard way to test this is by an experiment. In such an experiment, we distribute the patients into two groups. One group gets the new drug, we will call this group the treatment group, and the other group is the control group and gets an older drug as reference.

When doing an experiment as outlined in the previous paragraph, we could possibly get the following fictional data. We have two equal groups of 20 people where the control group gets the reference drug ("drug A") and the treatment group gets the new drug ("drug B"). In the control group of this fictional experiment, 10 out of 20 people recover and 10 do not, see figure 1a. In the other group 15 out of 20 people recover, see figure 1b.

(a) Control group (drug A) (b) Treatment group (drug B)

Figure 1: Drug Test

This is all the available data from this experiment. However, we could also have assigned every person to the other group, i.e. the control group could have been exposed to drug B and the treatment group to drug A. This is contrary to the fact of what we observed in the experiment, which is why we call this the counterfactual outcome. In the optimum case we would have these data as well as the data of the original experiment. When we have these data, the effect is easily calculated on an individual basis. In the end, this is what we want to achieve with medical research: acquire the best medical treatment for every person. As we outlined however, this is not possible, because we cannot observe all data. As a result, we can only draw conclusions from the data provided through the experiment as in figure 1.

It is obvious that we can only expose a specific patient to one drug. If we would expose the same patient to both drugs at the same time, we do not know which drug is responsible for the outcome. Moreover, the two drugs might be affecting each other. Whether that is positively or negatively does not matter, the data are useless.

In this fictional experiment we want to answer the question: "Which drug is better than the other?" At first, it might be obvious to conclude that drug B is better than drug A. The first thing people tend to rush into is asking the question whether a difference in success rate of 25% would be significant in this case Given the data from the experiment, this would be tempting to ask. However, as we stated earlier, we do not (and in most cases cannot) know what would have happened when the groups were exposed to the other drug. To illustrate the point that the conclusion that drug B is better than drug A is drawn too quickly, we introduce the fictional data when the groups were exposed to the other drug. In the control group, which is exposed to drug B this time, only 5 out of 20 people recover, see figure 2a. In the treatment group, this time exposed to drug A, everyone recovers, see figure 2b.

(a) Control group (drug B) (b) Treatment group (drug A)

Figure 2: Drug Test: Unobserved Data

When we have "all" the data of the 40 people in regards to the drugs, drawing a conclusion is easier and one can do this with more confidence, since there are no missing data. In both the control group and in the treatment group there are more people who recover from drug A than from drug B. These data are fictional however, but it illustrates that there is a problem when trying to draw conclusions from data acquired from an experiment: we do not know what the counterfactual outcome is. This leads us to the Fundamental Problem of Causal Inference, which will be defined and explained in the next section.

2 Fundamental Problem of Causal Inference

As we said in the previous section, it would be ideal to observe both values on the same unit. If we have both values of control and treatment on a unit, the causal effect is simply the difference between the treatment and the control. This estimated causal effect is on an individual basis. This is ideal in medical research for example, especially with the upcoming trend of personalized medicine, where the medication for a disease depends on the person.

Holland et al. (1985) state that the Fundamental Problem of Causal Inference lies in the fact that it is impossible to *observe* the values of control and treatment on the same unit and therefore it is impossible to *observe* the effect of the treatment on a unit.

In the Fundamental Problem we emphasize the word observe, as Holland et al. (1985) do. We do, because the impossibility to observe both the result when a unit is exposed to a cause (control for example) and the result when the same unit is exposed to another cause (treatment for example) is trivial in some cases, but less obvious in others. As we saw in the previous section we do not observe all data in a medical experiment. We only observe the effect of drug A on the patients assigned to the control group. Similarly, we only observe the effect of drug B on the patients assigned to the treatment group.

When we are dealing with another experiment this might be less obvious. For example, everyone experiences some problems with their computer. Some are caused by users, some are caused by bugs in software. For a fictional experiment, we want to determine what a certain button, part of some software program, does. One cause is that we click the button, another cause is that we do not. We click the button and the computer crashes. We repeat this a few times and every time we click the button, the computer crashes. We might be inclined to conclude that the button causes the computer to crash, but this is only because of certain assumptions that we can use to convince others of the correctness of the experiment. One of those could be that the computer is fine, except for that piece of software. If we knew that the computer crashes out of the blue, we might doubt that we observe both effects of the two causes. In the next section, we will explain this problem mathematically using Rubin's model.

3 Stochastic model

Before we explain the model for causal inference by Rubin, we will explain a model for associative analysis. This standard statistical model is widely used and provides a good introduction to the model for causal inference.

We use associative analysis to relate one data set to another. Where causal inference is all about causality, associative analysis is merely about describing the relationship between variables, in a descriptive sense. The difference is best made clear with a classic example: ice cream consumption is correlated to homicide rates, but there is no reason to expect this relationship to be causal.

The model we use for associative analysis is a standard statistical model which relates two variables concerning a finite population. We call the population U, consisting of elements or units u.[2] Most of the time we want to describe a relationship between two variables, we call them Y and A and both are functions on U. The value of these variables is a number, given by measurements on a unit u.

For example, we are interested in the relationship between eye color and hair length. We define A to be the color of the eye. We define Y to be the length of the hair. $A(u)$ gives information about eye color and $Y(u)$ about the hair length of unit u. We can look at the difference between the average hair length of people with blue eyes ($A(u) = 0$) and people with brown eyes ($A(u) = 1$). This describes a relationship between eye color and hair length. When we repeat this measurement after a certain time, the value of Y is probably different than the first measurement. The value of attribute A however is most likely the same.

This example exposed a subtle difference between Y and A. Formally, Y and A are of the same nature, because both are variables defined on U. However, we call A an attribute because its value does not change in time and is in that way different than Y.

In statistics we are interested in probabilities, distributions and expected values concerning variables on U. We use them to compare data sets. In this model, the probability of an event is nothing more than a proportion of units corresponding to that event in U. Conditional expected values are expected values of a subset of U, where this subset is defined by the condition in the expectation.

For example, the probability of a unit having a blue eye color, $P(A(u) = 0)$, is the proportion of the number of units in U with blue eyes divided by the total number of units in U. The conditional expected value of the hair length given a brown

[2] In more formal mathematical language, U is an outcome space where u is a possible outcome with a uniform distribution.

eye color, $E(Y \mid A(u) = 1)$, is the expected value or mean of the hair length of all units with brown eye color.

The majority of the information of this model is contained in the values of $Y(u)$ and $A(u)$ for all u in U. The joint distribution of Y and A is defined by $P(Y = y,\ A = a)$. $P(Y = y, A = a)$ is the proportion of u in U where $Y(u) = y$ and $A(u) = a$. An example is $P(Y = 185, A = 0)$ where Y, the height of a person, is 185cm and A, the eye color, is blue. We calculate the probability by taking the number of people that meet this condition and dividing that by the total number of people in U.

We are usually interested in the expected value given an attribute. For example, we want to know what the expected length is of people with a certain eye color. We call this the regression of Y on A and estimate it with $E(Y \mid A = a)$. One of the expressions to describe a relationship between to variables is association. We call this association α:

$$\alpha = E(Y \mid A = 1) - E(Y \mid A = 0). \tag{1}$$

This will explain some relationship between two data sets and associative analysis is descriptive in this sense. Causal inference uses equation (1) in some instances to estimate a causal effect (see section 4.2.1), but we need another model in order to draw those conclusions.

3.1 Rubin's model for Causal Inference

In the previous section we explained a model for associative analysis and discussed regression. In this section we explain Rubin's model for causal inference and point out what the problem is with this definition in section 3.2.

An experiment is not the only way to investigate causality, but it is the simplest method (Holland et al., 1985). It is important to note that whatever way we use to determine an effect, the effect of a cause is always relative to another cause. For example, "A causes B" almost always means that A causes B is relative to some other cause with the condition "not A". For example, when a drug heals you, it is the drug that causes you to be healthier than when you would not have taken the drug.

For simplicity, we use the language of experiments: treatment and control. Treatment is a cause and so is control, which is another cause. We determine the effect of the treatment relative to the control. It is important that there is always potential (but it might be hard or impossible in reality) for every unit to be exposed to both the treatment and the control. In some cases in reality, this is obvious. When you are sick for example, you can either take a drug (treatment) or not (control).

In terms of experiments, we assume that we work in a controlled experimental setting. That means that we are in control of what cause a unit is exposed to. There could be several causes, for example multiple drugs, or multiple categories of treatment.

Let us assume there are two levels of treatment: control (c) and treatment (t). Let S be a variable which indicates the level of treatment of a unit in U. In other words, $S(u) = t$ means that a unit u is exposed to t and likewise $S(u) = c$ means that a unit u is exposed to c. For every unit u in U with $S(u) = c$, $S(u)$ could potentially have been t. This also holds the other way around: for every unit u in U with $S(u) = t$, $S(u)$ could potentially have been c. Remember that we are in a controlled experiment and thus could have exposed the unit to the other cause. This is even true in uncontrolled studies, albeit out of control of the experiment.

Once set, $S(u)$ will not change. In that way, S is similar to A from the model for associative analysis (see section 3). A is an attribute of u and we can treat S as an attribute as well. There is one subtle difference however: S is forced upon a unit in a controlled experiment, while A is not. For example, $A(u)$ is the gender of a unit u, where $S(u)$ is the kind of drug where u is exposed to.

Analog to the model defined in section 3, we define a variable Y to be a function on U, given by a measurement on u. We are interested in whether an intervention has an effect on this variable. $Y_t(u)$ denotes the value of Y when unit u is exposed to t and Y_c denotes the value of Y when exposed to c. The model contains three variables, S, Y_t and Y_c. However, we only have two variables involved in the process of observation, S and Y_S.

In causal inference, time is important when it makes a difference *when* a unit u is exposed to a cause. When you are sick for example, the earlier you apply a treatment like a drug, the likelier it is that you recover. We divide the variables in two categories: pre-exposure variables and post-exposure variables. Pre-exposure variables are variables which are measured before exposure to a cause and post-exposure variables are variables which are measured after a cause. The post-exposure variables are thus variables that are possibly influenced by a specific cause, either t or c. Besides that, they can also be influenced by the point in time *when* a unit is exposed to a cause. This is nothing more than saying that causes have effect, which is the main principle of causal inference. We use the notation Y_c and Y_t for the post-exposure variables, where $Y_c(u)$ is the value of variable $Y(u)$ when u is exposed to c and $Y_t(u)$ is the value of variable $Y(u)$ when u is exposed to t.

We are interested in the effect of a cause. As we said earlier in this section, this is always relative to another cause. To measure the effect on a unit u, we take the

difference of $Y_t(u)$ and $Y_c(u)$. Holland et al. (1985) denote this with the algebraic difference:

$$Y_t(u) - Y_c(u). \tag{2}$$

This difference is what we call the causal effect of t on u. They stress that this effect is relative to the cause c. As we can see, the causal effect is defined for every specific unit u. Within this definition, there is a problem that Holland et al. (1985) call the "Fundamental Problem of Causal Inference". We discussed the problem in section 2. This problem will be discussed more mathematically in the next section.

3.2 Fundamental Problem of Causal Inference (mathematically)

With this model for causal inference, we can have another view of the Fundamental Problem. We state the Fundamental Problem once more and then explain the problem.

The Fundamental Problem of Causal Inference lies in the fact that it is impossible to *observe* the values of control and treatment on the same unit and therefore it is impossible to *observe* the effect of the treatment on a unit.

The impossibility to observe both $Y_t(u)$ and $Y_c(u)$ is trivial in some cases, but less obvious in others. For example, if the unit u is a computer and t means we double click on an mp3 file, c means we do not and Y indicates the computer is playing music after c or t, then we might believe that we know the values of both $Y_c(u)$ and $Y_t(u)$ by simply double clicking the file. However, this is only because of a certain assumption that we can use to convince ourselves and others of the correctness of the experiment. If for example the computer was randomly playing the music file because of some internal error, we might doubt that we know the values of both $Y_c(u)$ and $Y_t(u)$.

Another example is the medical experiment from section 1. In that case, it is obvious that we cannot observe both Y_c and Y_t. We can only apply either drug A or drug B to a patient. If we exposed a patient to both of them, the results would be unusable, because we do not know which drug is responsible for the outcome.

The implicit problem that follows from the Fundamental Problem is that causal inference is impossible. Holland et al. (1985) claim that we should not jump to this conclusion too quickly. By saying we cannot observe both values, we do not mean to say we cannot derive information about them. We will explain some solutions to this Fundamental Problem in the next section.

4 Solutions to the Fundamental Problem

There are two types of solutions for the Fundamental Problem of Causal Inference. Holland et al. (1985) call one type the scientific solution and the other type the statistical solution. We will discuss both in this section and give examples when they could be applicable. An important aspect of these solutions is that the applicability depends on the extent to which the assumptions are correct, but we will return to this issue within each solution.

4.1 Scientific Solution

The scientific solution is to make assumptions about homogeneity and invariance. Invariance is the property that the experiment is the same at a specific point of time in the future (see section 4.1.1). Homogeneity is the assumption that the units are identical in all the relevant aspects (see section 4.1.2). All of these assumptions are non-testable. However, one can convince themselves and others that the assumption is correct. By doing so carefully, one can claim the correctness of an assumption, but never prove it.

4.1.1 Temporal stability and causal transience

A way to apply the scientific solution is to assume that:

(a) The value of $Y_c(u)$ does not change when u is being exposed to c and $Y(u)$ is being measured.

(b) The value of $Y_t(u)$ does not change when u is being exposed to t after also being exposed to c as in (a).

It is simple to calculate the causal effect when both of these assumptions are plausible. Indeed, we can first expose u to c and then to t, while measuring Y after each exposure. The first assumption is the temporal stability, from which follows that the response to a cause is constant. The second assumption, the causal transience, claims that the effect of c on u does not change the later measured $Y_t(u)$.

For example, we can setup an experiment where we test a new version of aspirin. We define U to be a "super population", consisting of people where the state 'needs aspirin' may occur. This way, u is a person who sometimes needs an aspirin. We denote the control and treatment variables as follows: c is applying the old drug and t is applying the new drug. Y indicates the extent in which u gets better. In this example, it is plausible to assume both (a) and (b). One can convince others that these assumptions hold, by saying that the old aspirin will leave the body of u in a specific amount of time. You cannot prove this however. The same applies to

the other assumption, the temporal stability, which might be true and arguable, but cannot be proved. If plausible, the causal effect of t can be easily calculated, we can first expose u to c and then at a later moment to t, while measuring Y after each exposure. The exposure to t is when u reaches the state 'needs aspirin' again. Then $Y_t(u) - Y_c(u)$ is the causal effect, on an individual basis.

4.1.2 Homogeneity in units

A second way to bypass the Fundamental Problem of Causal Inference is to assume $Y_t(u_1) = Y_t(u_2)$ and $Y_c(u_1) = Y_c(u_2)$ for each two units $u_1,\ u_2 \in U$. Holland et al. (1985) call this assumption the homogeneity in units. The causal effect of t is the value of $Y_t(u_1) - Y_c(u_1) = Y_t(u_1) - Y_c(u_2)$. In other words, when dealing with two related units, one can estimate the effect by applying one cause to one unit and the other cause to the other unit.

As with the other assumptions, this one cannot be proven. One can convince others of the plausibility however. For example, when doing an experiment to test a new drug, you can take twins and apply one treatment to one of them and the other treatment to the other. This is possible if you can assume that the twins are identical in all the relevant aspects in regards to the drug.

Another example where this assumption can hold, is when we have cloned units. The last few years we are increasingly able to clone and we can use this in medical research to research the effect of certain interventions. The reason why we can do this is because of the assumption of homogeneity.

The assumption of homogeneity implies a constant effect, in other words $Y_t(u_1) - Y_c(u_1) = Y_t(u_2) - Y_c(u_2)$. The effect is then constant across all units in U. We will elaborate on this relation between the homogeneity and the constant effect in section 4.2.2.1.

4.2 Statistical solution

Besides the scientific solutions, there are also statistical solutions. While the scientific solution uses assumptions about homogeneity and invariance, the statistical solutions uses characteristics of a population. We will start with explaining the average causal effect.

4.2.1 Average causal effect

Up to this point, we were mostly interested in the causal effect on a particular unit. However, we can also look at the average causal effect on a set of units. That also corresponds to the way we conduct most of our experiments. We inspect a group

of people to estimate an average effect. The way we conducted our experiments in the past can indeed be correct, as we will show below.

The assumption we use is that the population U consists of a large set of units. We conduct an experiment and observe pairs of the variables (S, Y_S). Recall that this is the cause to which the unit is exposed and the result of that cause as measured by Y. The average causal effect T can now be estimated from $E(Y_t)$ and $E(Y_c)$. The observed data gives information about $E(Y_S \mid S = t) = E(Y_t \mid S = t)$ and $E(Y_S \mid S = c) = E(Y_c \mid S = c)$, since that is the way we set up the experiment.

It is important to note that $E(Y_t)$ and $E(Y_t \mid S = t)$ are in general different. Indeed, not every unit u is exposed to t. The same holds for $E(Y_c)$ and $E(Y_c \mid S = c)$. There is however an assumption where they are equal, which is again (luckily) common in experiments. The assumption is that the units in U were randomly assigned to either the control group or treatment group. Then we can estimate the average causal effect T:

$$T = E(Y_t) - E(Y_c) = E(Y_t \mid S = t) - E(Y_c \mid S = c). \tag{3}$$

Both terms on the right hand side are known when conducing an experiment and as such, the average causal effect can be calculated. Recall that the associative parameter in 1 was defined as $\alpha = E(Y_t \mid A = 1) - E(Y_c \mid A = 0)$. When we change A to S, we have $\alpha = E(Y_t \mid S = t) - E(Y_c \mid S = c)$. When Y and S are independent, that is the treatment is assigned randomly, $\alpha = T$.

Relative to the other solutions for the Fundamental Problem of Causal Inference, this one is fairly simple to achieve and to convince others of its correctness. There is a big disadvantage though: the average causal effect does not indicate whether any particular individual would be positively or negatively affected by the treatment. In medical research, this is unfortunate to say the least. Some parts of the population might be harmed by a drug, even though the average causal effect is positive. This is at odds with the intentions of personalized medicine, where the treatment is adapted to the patient.

4.2.2 Constant effect

The average causal effect T is an average and Holland et al. (1985) state that it likewise has all the benefits and disadvantages of an average. Like we said in section 4.2.1, the average causal effect does not indicate whether any particular individual would be positively or negatively affected by the treatment. The assumption of a constant effect solves this problem, at the cost of another assumption.

As the name implies, the assumption is that the causal effect is constant among the population. Every unit is equally affected by exposure to the treatment. In other words, $T = Y_t(u) - Y_c(u)$ for all u in U and thus constant. The assumption of a constant effect allows the value of an average causal effect to be relevant for every unit and thus allows to draw causal inferences based on T on individual units.

4.2.2.1 Relation to the homogeneity assumption

Holland et al. (1985) remark that the assumption of a constant effect is implied by the homogeneity. In other words, when $Y_t(u_1) = Y_t(u_2)$ and $Y_c(u_1) = Y_c(u_2)$ then it is trivial that $Y_t(u_1) - Y_c(u_1) = Y_t(u_1) - Y_c(u_2)$. Thus we can see the assumption of a constant effect as a weakening of the homogeneity assumption.

The situations where the assumption of a constant effect is correct or at least realistic are rare. The treatment has to have the same effect on all units in U. When dealing with patients in medical research, this is debatable. People are all different and could potentially react different to a treatment. However, there are some cases where the assumption of a constant effect can hold. All cases where there is unit homogeneity there is also a constant effect, as we showed in the previous paragraph. Because of that, we can test a drug on twins, where we assume and check, if possible, that they are similar in all relevant aspects of the experiment. For example, they have the same DNA, same history on alcohol consumption and both are not smoking. The list can go on indefinitely with characteristics of the twins which could be relevant for the research. When we do have two "identical" persons, we can assume there is a constant effect because of the homogeneity assumption.

Another example with a constant effect, where there is no homogeneity, is hard to find. We can even ask the question if there are any of these instances. It does not help that the assumption of a constant effect is hard to prove. We can assume this in a fictional experiment. We want to investigate how the amount of sunshine affects the growth of a certain type of plant. We use U for the units, in this case the plants. The control group gets 0 hours of sunshine, which we denote with c. The treatment group gets exposed to t, 4 hours of daily sunshine. Y is the height measured after one week of the plant. When we assume there is a constant effect, we can calculate it with $T = Y_t(u) - Y_c(u)$ and this holds for all plants by the assumption of a constant effect.

In this case, it might be plausible that the effect is constant. It does not matter what plant you have, they could all benefit the same amount from sunshine hours and all die when not exposed to sunshine. In this fictional experiment, it could also be that the plants are similar in all relevant aspects and thus comply with the homogeneity assumption. This could easily be the case: all plants die when not exposed to sunshine.

4.2.2.2 Constant effect over time

Many experiments have taken place in the past. Whenever we wanted to research something, we used an experiment and still do. However, the results of the past might not correspond with the results when the experiment was conducted again at a later moment. For example, smoking could have possibly decreased over the course of the last 20 years (NHS Digital, 2017). This could possibly have influenced the results of experiments in health care, to give an example. We could extend the assumption of constant effect for some cases where we are sure that the effect is also constant in time. This would mean that the experiment only has to be conducted once to estimate the effect, which is constant across units in time. This assumption where an effect is constant in time is however not provable. This thus makes it hard to convince others of its correctness, which is what we want to defend our results. The reason that one cannot prove this is because even in the future, with unpredictable changes, the effect has to be constant.

4.2.3 Causal inference in observational studies

There are many studies where we would like to know what the effect is of something, but cannot study this with an experiment. This could be because of limitations originating from the subject, or ethical limitations. An observational study is done in these cases. With an observational study, people are not randomly divided in a control group and treatment group, but they "choose" their group by their behavior. For example, when we would be studying the causal effect of alcohol consumption on life expectancy, we cannot force people to drink a certain amount of alcohol, they do it themselves.

Holland et al. (1985) say that an important idea is that the pre-exposure variables can be used to replace the independence assumption with the weaker conditional independence assumptions. According toHolland et al. (1985), this idea stems from Rubin (1974, 1977, 1978), Rosenbaum (1984a, 1984b, 1984c), Rosenbaum and Rubin (1983a, 1983b, 1984a, 1984b, 1985a, 1985b), Holland and Rubin (1983).

The extent to which the results from observational studies are true is hard to

estimate. There are too many variables to draw conclusions from one study, for example: how did you question people, is there a selection bias and are there confounding variables? So when can we use the results of an observational study? Wasserman (2004) said that results from observational studies are credible when:

1. The results of the study correspond to the results of earlier studies.
2. Every study is controlled by plausible confounding variables.
3. There is a plausible scientific explanation of the existence of the deduced causal relationship.

We will apply this knowledge in an example in the next section.

5 An illustration: Alcohol Consumption

We apply the knowledge from the last section to an example. We use the assumptions for causal inference to illustrate that they are easily overlooked when drawing causal relations.

An extensive research was done by Wood et al. (2018) in April 2018 about risk thresholds for alcohol consumption. Wood et al. (2018) studied individual-participant data from nearly 600 000 current drinkers without previous cardiovascular disease. The data were collected from three large-scale data sources in 19 high income countries. The objective was to find a risk threshold for alcohol consumption and death (all causes) and sub types of cardiovascular diseases. The risk threshold was the amount of alcohol with the lowest risk for all-cause mortality and cardiovascular disease.

5.1 Data collecting

Wood et al. (2018) only looked at people who do drink, to prevent the following:

1. Former drinkers could have stopped drinking because of health problems. This would result in a bias in the non-drinkers, because the unhealthy former drinkers are included in the same group.
2. Residual confounding; the distortion that remains after controlling for the confounding variables through the design of a study. All observational studies are by definition experiencing distortion because of residual confounding. The distortion caused by the people who do not drink was better to avoid.
3. People that have never drunk alcohol can systematically differ from drinkers in immeasurable or difficult to measure ways, but are important in regard to the research.

Wood et al. (2018) collected data from other studies, where the participant needed to keep track of information about the following for at least a year: alcohol consumption and status, as well as age, history of diabetes and smoking. In addition, the participant should not have had cardiovascular diseases in the past.

The data was then categorized in 8 categories. The alcohol consumption was therefore converted to consumption in grams to have a standard scale across all data. The categories were made on the basis of this new scale: $(0, 25]$, $(25, 50]$, $(50, 75]$, $(75, 100]$, $(100, 150]$, $(150, 250]$, $(250, 350]$ and $(350, \infty)$ gram per week.

5.2 Findings

In the 599 912 current drinkers included in the analysis, Wood et al. (2018) recorded 40 310 deaths and 39 018 incident cardiovascular disease events during 5.4 million person-years of follow-up. These data made the following finding: "In comparison to those who reported drinking $(0, 100]$ gram per week, those who reported drinking $(100, 200]$ gram per week, $(200, 350]$ gram per week, or $(350, \infty)$ gram per week had a lower life expectancy at age 40 years of approximately 6 months, 12 years, or 45 years, respectively." (Wood et al., 2018)

5.3 Media

Several media picked up this news, each with a different story. The paper itself did not draw a conclusion, neither did it make a causal relationship. The news media however, did draw conclusions. We will look into a few of them and explain why their conclusions were wrong and how it could have been right when the original study was done differently.

5.3.1 AD

AD (Algemeen Dagblad), a Dutch newspaper, wrote the following on this topic, in response to this research: "An extra glass of alcohol can shorten your life with 30 minutes. Those who want to become as old as possible, are better off not drinking any alcohol." (Buitenlandredactie, 2018)

The article in the newspaper assumes implicitly that the participants are assigned randomly to a level of alcohol consumption and that the effect of alcohol consumption is the same on every individual. If both assumptions were true, there would still be one minor problem: the research does not say anything about those who do not drink alcohol and thus the conclusion that you are better off not drinking is not correct.

Besides that, a lot of information, like the age of 40 and when extra is extra, is further down the article of the newspaper, which means that it is less important in the eyes of the journalist. It is also important to note that the original research article does not draw causal conclusions and that this is all "derived" from the research article.

5.3.2 The Guardian

This mistake was not only made in The Netherlands, but also in English newspapers, like The Guardian. This newspaper also headlines with the fact that one extra glass of alcohol a day will shorten your life by 30 minutes (Boseley, 2018). In the rest of the article, this saying gets nuanced. However, "on average" gets left out a lot, which implicitly assumes that the effect of alcohol consumption is the same on every individual.

5.3.3 De Volkskrant

De Volkskrant, another Dutch newspaper, published something along the lines of what the other newspapers did. However, they retracted this in a revision. The conclusion that every glass was shortening your life by 30 minutes was "not in the

study". Half a year, one and a half year earlier or 30 minutes earlier dead per glass is "juggling with statistics". (Sahadat, 2018)

5.4 Carefulness

As showed in this example, a mistake regarding causal inference is easily made. Therefore, it is advisable to be aware of these pitfalls when investigating something or writing something about a research report. The most important thing to watch out for is being too eager to conclude a causal effect.

6 Summary

In summary, this thesis gave an introduction to causal inference. First of all, we showed an example where we elucidated the difference between treatment and control. In medical research, there is always a control group and a treatment group. The control group gets one drug and the treatment group another one, all with the goal to find out which drug is better.

We saw that a specific patient can only be exposed to one of the two drugs and not (simultaneously) to both. We are however interested in what the other drug would have done on the same person. This is usually not possible. That is one of the reasons that incorrect conclusions are drawn from experiments due to incomplete data. The impossibility to observe the effect of both drugs on the same person is in essence the Fundamental Problem of Causal Inference.

The Fundamental Problem of Causal Inference lies in the fact that it is impossible to *observe* the values of control and treatment on the same unit and therefore impossible to *observe* the effect of the treatment on a unit. (Holland et al., 1985) The emphasis is on the word observe, because it is impossible to observe both the result when a unit is exposed to a cause and the result when a unit is exposed to another cause.

We used a standard statistical model to later introduce the model for causal inference. We use associative analysis to relate one data set to another. In order to do so, we define a population U, a variable A and an attribute Y. These are formally of the same nature, but A will not change in time. The majority of the information of this model is contained in the values of $Y(u)$ and $A(u)$.

The model we used for causal inference is Rubin's model. We assumed that there are two levels of treatment: control and treatment. Both are causes and we determine an effect of a cause always relative to another cause. In this case that means treatment relative to control, where control could also be "not treated". It was important to think about the fact that the cause to which a unit was exposed always could have been different, even in observational studies.

We stated that time did not matter in associative analysis as much as it did in causal inference. In causal inference, time is important when it makes a difference *when* a unit u is exposed to a cause. To make a difference between variables that are influenced by time, we divided the variables in two categories: pre-exposure and post-exposure variables. The post-exposure variables are those variables that are influenced by a specific cause and possibly influenced by the time at which a unit is exposed to a cause.

We explained this model because we are interested in the effect of a cause. To

measure the effect on a unit u, we calculate $Y_t(u) - Y_c(u)$. This however gave the problem we saw earlier: the Fundamental Problem of Causal Inference. The implicit problem that follows from this is that causal inference is impossible. However, we can derive information about both values under certain assumptions.

We discussed a range of assumption to make it possible to estimate the causal effect, or the weaker average causal effect. None of them are provable, the best one can do is convince themselves and others of its correctness. We divided them into two categories: the scientific solutions and statistical solutions. Temporal stability and causal transience was one of the scientific solutions, which assumes that the effect fades away over time and thus makes it possible to measure the effect of both causes at a different time.

Another solution, also scientific, was the homogeneity in units. This was the assumption that the effect of a cause is the same across two units because the units are similar in all relevant aspects. In this case, the effect equals $Y_t(u_1) - Y_c(u_1) = Y_t(u_1) - Y_c(u_2)$ and both can be measured.

Then we discussed the statistical solutions, where we used characteristics of a population to calculate the (average) causal effect. The first one was the average causal effect. Instead of looking at the effect on an individual basis, we did look at it on a population basis. We randomly assign treatment and control to calculate the average causal effect $T = E(Y_t) - E(Y_c) = E(Y_t \mid S = t) - E(Y_c \mid S = c)$. This is the simplest to achieve of all assumptions, by correctly setting up an experiment, but we lose some detail. We can no longer make any statements on an individual basis with the average causal effect.

We dealt with this unfortunate side effect by assuming that the effect is constant across all units. This is a weakening of the assumption of unit homogeneity and often this assumption is better to argue. When we introduced time we obtained an even stronger result, the conclusions regarding the effect are more powerful, but the assumption became impossible to prove.

The last assumption is a weakening of the average causal effect to use it in observational studies. In observational studies it is impossible to assign people randomly to be exposed to a cause. However, there are three points that make the results of an observational studie credible (Wasserman, 2004):

1. The results of the study correspond to the results of earlier studies.
2. Every study is controlled by plausible confounding variables.
3. There is a plausible scientific explanation of the existence of this causal relationship.

Finally, we have taken a look at the media and their story about a recent research regarding alcohol consumption. It went wrong in translating an associative conclusion into a causal relationship. However, one of the newspapers did a revision and came back to their original article. They all however misinformed their readers initially and need to be more careful when drawing causal relations from studies. The most important thing to watch out for is being too eager to conclude a causal effect.

7 Discussion

A recommendation for future papers regarding causal inference is a more mathematical way of describing the model for causal inference, i.e. by setting up a probability space and work from there.

For future research, it would be interesting to have a qualitative study where mathematicians reach out to journalists to find out their way of thinking and their motives to do so. That way there might be a point in time where the reader is no longer misinformed by causal relations based on associative studies. Moreover, every causal relation then drawn will be based on a correctly executed study and thus more reliable.

References

Boseley, S. (2018, April). Extra glass of wine a day 'will shorten your life by 30 minutes'. *The Guardian*. Retrieved from `https://www.theguardian.com/science/2018/apr/12/one-extra-glass-of-wine-will-shorten-your-life-by-30-minutes`

Buitenlandredactie. (2018, April). Een extra glas alcohol kan je leven met 30 minuten verkorten. *AD*. Retrieved from `https://www.ad.nl/buitenland/enlsquo-een-extra-glas-alcohol-kan-je-leven-met-30-minuten-verkorten~a0ff48af/`

Holland, P. W., Glymour, C., & Granger, C. (1985). Statistics and causal inference.

Holland, P. W., & Rubin, D. B. (1983). *"on lord's paradox" in wainer and messick prinicpals of mordern psychological measurement*. Hillsdale, N.J..

Huff, D. (1954). *How to lie with statistics*. New York: W. W. Norton & Company.

NHS Digital. (2017). *Statistics on smoking, england 2017*.

Rosenbaum, P. R. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignmen. *Journal of the American Statistical Association*, *79*, 41–48.

Rosenbaum, P. R. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society*, *147*, 656–666.

Rosenbaum, P. R. (1984c). Conditional per.mutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, *79*, 565–574.

Rosenbaum, P. R., & Rubin, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*, *45*, 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1984a). Discussion of pratt and schlaifer "on the nature and discovery of structure.". *Journal of the American Statistical Association*, *79*, 26–28.

Rosenbaum, P. R., & Rubin, D. B. (1984b). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38.

Rosenbaum, P. R., & Rubin, D. B. (1985b). The bias due to incomplete matching. *Biometrics*, *41*, 103–116.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Sahadat, I. (2018, April). Nog één keer over dat ene glaasje alcohol. *De Volkskrant*. Retrieved from `https://www.volkskrant.nl/nieuws-achtergrond/nog-een-keer-over-dat-ene-glaasje-alcohol~bf4c005a6/`

Wasserman, L. (2004). *All of statistics*. New York: Springer.

Wood, A. M., Kaptoge, S., Butterworth, A. S., Willeit, P., Warnakula, S., Bolton, T., ... Danesh, P. J. (2018, April). Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599912 current drinkers in 83 prospective studies. *The Lancet*, *391*(10129), 1513–1523.