Jasper Henze
Studiengang: Angewandte Informatik Bsc.
Matrikelnummer: 3365720
Email: j.henze@stud.uni-heidelberg.de
Seminar: How to Lie with Statistics
Prof. Dr. Ulrich Köthe

Reintegrating Causality into Statistics OR In Search for a Formal Language of Causality

"Correlation does not equal Causation". Almost everyone who has worked with statistics has heard this sentence at some point in time. And it isn't hard to find every day examples that underline the importance of this truism, e.g. a statistical study might show that people who are in a hospital are more likely to die in a given time interval than people who are not in a hospital. Assuming that the correlation between being in a hospital and death rate would be a causal relationship would indeed be fatal, since the conclusion would be that being in a hospital makes people more likely to die and therefore it would only serve the public health to remove hospitals alltogether.

And yet this observation proved detrimental for the field of statistics, because the founding fathers of this new science did not take the next step of uncovering the more hidden mechanisms that were at work in the example above, but instead decided to villify causation and declare that statistics can only be used to observe correlations and nothing beyond that.

In fact famous statistician Karl Pearson conducted an experiment similar to the one mentioned above where it was obvious that the measured correlation was not of causal nature. He obtained data on the measurements of the length and breadth of a large number of skulls (page 70 *The Book of Why*). He observed that there was no correlation between skull length and breadth if looking at the skulls of only one gender. Yet when he combined the data of both genders a significant correlation between the two measurements could be observed. Pearson called this phenomenon "spurious correlation" and took it as proof that one cannot "persist in looking upon all correlations as cause and effect" (page 71).

Though this assessment was correct his conclusion to completely turn his back on causality was a grave oversight. If we look at the data again, we can explain this spurious correlation by unraveling the causal mechanisms that were working in the background. The reason that Pearson got a

correlation between skull length and breadth was that there was a third factor that affected both measurements but differed in the two populations. Since women tend to have smaller heads than men they also tend to have smaller head breadth and length. So now suddenly, a small skull length would indicate that the skull is more likely to belong to a female, and therefore also more likely to have a smaller breadth.

The previous example with the correlation between hospitals and death rates follows the same pattern. In this case the third factor would be illness which affects the likelihood of being in a hospital as well as death rate. If one would account for this factor by grouping test subjects depending on whether they are ill or not, and on the severity of their sickness (similar to what Pearson did when he looked at the skull measurements of only one gender), one would hopefully find that hospitals do not increase death rates but overall lower them. So we would still have a correlation, but this time with the opposite effect and one that is also identical with what we would have predicted as the causal relationship. (After all hospitals are built in order to prevent people from dying to diseases/wounds).

And even though this explanation seems quite logical and intuitive, a few decades ago it would have been received with harsh criticism for using causality in a statistical problem. Recently however, a form of statistics that uses the concept of causality has been developed and popularised. One of the scientists leading this "causal revolution", as he coined it himself, is Judea Pearl who wrote the *Book of Why* in which he tries to convince the reader of the necessity of using causal reasoning in statistics and in which he presents a formal language with which causal problems can be formulated.

Pearl states that there are at least three different levels to causal reasoning (page 27). The first level is that of association. Here one is limited to observing correlation between variables and acting accordingly. This is the level that statistics restrict themselves to if they completely disregard causality and only work with observational data. Spurned by the idea that everything can be gained from observed data and that by creating models, the scientists would not be able to get over their prejudices about how they imagine things to be, these statisticians would end up in a trap of their own devise, where they would not have an alternative to the above conclusion of the hospital example, that hospitals lead to a higher death rate, other than saying that correlation does not imply causation and therefore ultimately does not imply anything, rendering any statistical studies practically useless. This might seem unreal and yet the debate over the health effects of smoking in the 50's and 60's shows that statistics was in exactly that situation and how helpless they were at

giving an answer to the important question asked.

Another example of entities that are restricted to this first level of causation are current machine learning programs. If one would want to support the claim that it really is all in the data, deep learning problems woul probably be the best way to do so. After all the succes of popular deepmind programs by google speaks for itself. Yet while they are unbeatably strong at tasks with complete information such as winning chess or go matches they fail at most other tasks. One game where deep learning's limits became quite apparent was the video game Starcraft. In Starcraft a lot of information is hidden from the player as he can only see wherever he controls units or buildings on the map. This introduces uncertainty into strategies and allows uncountable risky strategies, where a player tries to disguise his gameplan by hiding key structures while pretending to be following a completely different strategy. While the AI eventually became very good (though not perfect) at playing standard games in which the opponent did not employ bluffing but instead followed a very efficient build order, even beating the reigning world champion at such games, it struggled heavily when facing unorthodox strategies. In addition to that, Alphastar's associative roots sometimes became very glaring when it copied certain strategic ideas from humans since they were associated with a high success rate, yet it implented these ideas into its own builds in ways that made it very clear that it did not understand why they work. One such example would be the construction of a wall at the entrance of one's base with a hole in it that is guarded by one single unit and functions as a sort of door. Since every succesful player was using this strategy, Alphastar copied this behaviour, yet it would build inefficient walls with sometimes even more than one hole, demonstrating its complete lack of understanding of the causal mechanisms that humans use to make such strategic decisions.

One more flaw of deep learning is its intransparency. The developer's of an AI can't predict the results of their creations and instead often have to use the age old method of trial and error. On the other hand the reason that the AI makes its decisions also remains a mystery. Asking a human why he made a specific chess move will yield an answer that explains the state of the board, which figures threaten which, and how the player hopes to gain an advantage by making this specific move, and the future possible pathways that could result from this move. The deep learning AI can only really give as an answer "Because it is the best move to do". Even though this can still be insightful to chess players as this can allow them to now in turn to justify that move and try to find the reason as to why it is good; the human answer, even though the move is not necessarily as optimal as the AI's, can give much more insight into the game to someone trying to understand a game of chess than the AI's answer would.

The improvement of AI towards a strong AI that is based on causal language is one of the main motivations for Pearl when it comes to his research in Causality. In order for an AI to be transparent and be able to discuss the reasons for its actions with humans it must ascend past the level of association and instead be able to answer the "Why-Question".

The second level of causal thinking is that of intervention. On this level the statistician is no longer a mere passive observer of the world that he lives in, but instead an active subject that can manipulate the conditions under which an experiment takes place in a way that might help him uncover the effect that he was looking for. It encompasses the ability to ask the question "What happens if I do something specific?". A good example of individuals on this level would be babys and small childs, who are constantly trying out new things and always testing out new waters. There is probably no better way to truly learn about the disadvantages of touching a hotplate than doing it. Though this example already hints that this method can be rather risky and sometimes unsuitable due to practical or ethical reasons.

For statistics intervention is a very strong tool when it comes to discovering causal relationships. If one would apply this principle to the previously described hospital example, this would mean that instead of just looking at observational data of a population in its "natural state", one would create an artificial environment where whether a person is in a hospital or not would no longer be decided by external factors such as illness, but instead every person would randomly be assigned to either go to the hospital or not without considering any of the usual factors. Now illness would no longer confound the effect of being in a hospital on death and more so any other possible confounders that we didn't think about are also automatically eliminated, since going to a hospital or not now only depends on one independently determined random number. Thusly now any correlation would indeed be causation and we would see the true effect of hospitals on death rate.

Now this setup would be pretty unrealistic, as healthy people usually wouldn't go to the hospital (unless they are working there). A more appropriate setup for answering this would be assigning test subjects randomly to one of two groups where one group can't go to hospitals and the other group can, but only if they would do so under normal circumstances. The first group is the treatment group that would give an answer to how having no hospitals would affect overall death rate, and the second group would be the control group that gives us data of the status quo that can be used for comparison.

This method is called the RCT (short for Randomized Controlled Trial) and is one of the most popular methods in statistics since it guarantees the removal of the most common enemy of any

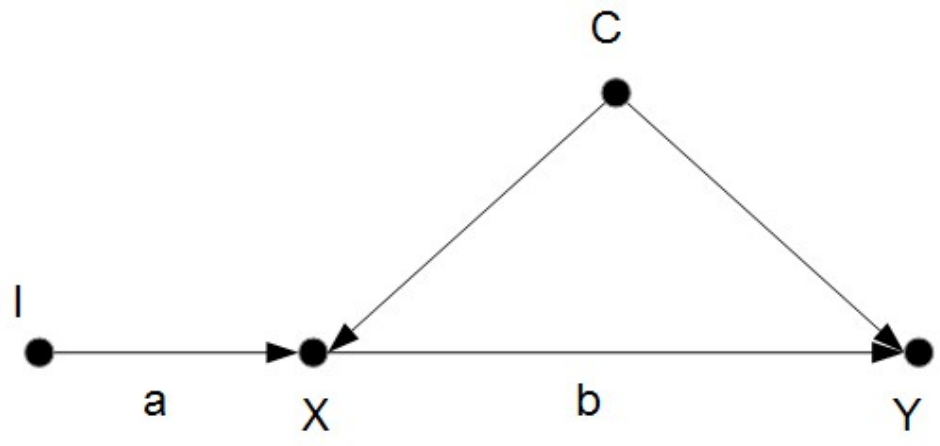
study: unknown or unobservable confounders.

The concept of the RCT was first described by R.A. Fisher in the 1920's and has since then been accepted by all statisticians as a way of discovering a causal effect (page 143). Now with that being said one might think the RCT solves all of statistics problems with causality but as the example above shows us, there are multiple drawbacks that come with the RCT. If one however can overcome all of these difficulties one will have an extremely high probability of discovering a real causal effect.

Ethical concerns are a very common tripwire for RCTs. In fact looking at the above hospital example, one can find several ethical problems when trying to set up a RCT experiment. Denying sick people access to a hospital, even if it is for the sake of scientific progress, can only be described as morally questionable. Another problem is that, in contrast to observational data which is often much more easily collected and potentially already available, setting up experiments that have sufficiently large sample sizes usually requires a lot of time, money and willing test subjects. Luckily however there are multiple ways of transforming questions that would usually require intervention in such a way that they can be answered with only observational data.

Pearl dedicates an entire chapter to describing these methods. These include but are not limited to the convential checking and adjusting for confounders. For cases in which confounders cannot be observed or are unknown the front-door criterion or an instrumental variable can be used to uncover the causal effect unspoiled by confounders if specific conditions are met.

The front-door criterion cann be applied whenever the effect in question is mediated by another effect. That means while we are interested in the effect of the treatment variable X on the outcome Y, the causal effect is not direct but rather X causes another effect M, the so-called mediator, which in turn causes Y. Now if we want to eliminate the effect of a confounder of X and Y(it is important that M is not affected by this confounder) that we can't observe and therefore can't adjust for, we can do so by measuring the effect of X on M and the effect of M on Y and multiplying the two. This



of course assumes that X has no direct effect on Y and only works through M and that the effects are linear.

Instrumental variables are another strategy that can eliminate unobservables and that applies when the roles between the treatment variable and the mediator in the previous example are swapped. That means that the treatment variable X affects Y directly and X and Y are confounded by a common unobservable variable C. But this time X is also the mediator when it comes to the effect that the instrumental variable I has on Y, that means I affects Y through X and, for this strategy to work, must also be unaffected by the unobservable confounder C.

The above diagram shows these relationships. The dots represent the described variables and any arrow that leads from one dot to another one symbolises the causal effect of that variable on the other one, with a and b denoting the strength of these effects. These simple diagrams are called path diagrams, as they outline causal pathways, and are a very commonly used method by Pearl for symbolising causal relationships.

In this case we are interested in measuring the value of the path coefficient b since this would tell us the causal effect of X on Y. However, due to the unobservable variable C affecting both X and Y, the observed correlation between X and Y will not be the same as the causal effect, no matter how big and exact our study is, as it will always be tainted by a systematic error.

If we can however measure the effect of the instrumental variable I on the treatment variable X (which would yield the value for the path coefficient a), as well as the effect of I on the outcome Y (equal to the product of path coefficients a and b), we could gain the sought after value by dividing the latter by the former effect: (Effect of I on Y)/(Effect of X on Y) = ab/a = b.

In *The Book of Why* Pearl underlines the practical significance of this technique with a historical example, where a study that used the above described strategy could have saved thousands of lives from cholera, if the results had gained more publicity.

In the mid 19th century, England was plagued by the cholera epidemic. Back then it was generally assumed that the disease was caused by unclean air which seemed to be supported by the fact that it was stronger in poorer districts where sanitation was worse. The physician John Snow however theorised that the disease was transmitted through either food or water, as the first symptoms occured in the intestines. Now in general food as well as water were cleaner in the wealthier districts of London than in the poor ones, which meant he had trouble setting himself apart from the original theory. However as luck willed it, there were two different large water supply companies back then, which supplied the same districts but drew their water from different areas of the river Thames. One, the Southwark and Vauxhall Company, drew their water from a spot that was

downstream from the point where the sewers entered the river, while the other, the Lambeth Company, had recently moved their water intake to be upstream from the sewers.
This setup can now be inserted into the above described path diagram. The instrumental variable I is the water company, the water pollution is the treatment variable X, and suffering from Cholera is the outcome Y, while C represents hard to measure factors such as hygiene and poverty which could affect both X and Y. Snow already knew the impact that water company had on water purity, simply by knowing that one used sewer contaminated water and the other one didn't. Following our model, all he had to do in order to find the effect of water purity on cholera, was measuring the effect of water company on cholera affliction. And that he did. He went from household to household and inquired to know which water company they were served by as well as how many cases of cholera they had. His research yielded that those households that were served by the Southwark and Vauxhall Company generally had a higher number of cholera cases than those served by the Lambeth Company, thusly proving his hypothesis that the disease was transmitted though food or water intake, in this case the pathogen left the body of the sick by the extreme diarrhea they suffered from and found their way through the sewers into the river back into the households of London and into the stomachs of their new victims.
It is worth noting that the only reason the water companies could be used as an instrumental variable was that they did not factor in poverty or hygiene when deciding which households to serve and therefore were shielded from these confounders. In fact Snow himself stated that "no experiment could have been devised which would more thoroughly test the effect of water supply on cholera ..." (page 247). And it does indeed seem as if the water companies had set up a RCT, even it was unintentionally, allowing Snow to answer rung two questions by only collecting observational data.
And yet even though it almost seemed like all the stars aligned and Snow found an undeniable proof as to what caused cholera and how to prevent it, his results gained no publicity and did not lead to actions by the water companies who could have prevented thousands of death by cholera by taking action and improving the purity of their water.

While the previously described strategies allow fitting rung two questions to observational data if specific conditions are met, they fall flat in all cases where they aren't met. This begs the question, could one devise a tool that can transform any such question into one that can be answered with only observational data? As a matter of fact such a tool already exists and was created by Pearl himself. The DO-Calculus is built upon three axioms and promises that every question about the

true causal effect of a variable upon another can be answered without having to conduct a RCT but only requiring a strong enough model as well as observational data.

To understand the axioms of the DO-Calculus, one first has to be familiarised with the notation of statistical probabilities and the differentiation between observed probabilities and interventional probabilities. For example P(Fire | Smoke) notes the probability of there being fire if one can see smoke, while P(Fire | do(Smoke)) describes the probability of fire if one is to create smoke. While P(Fire | Smoke) can be estimated to be fairly high, since fire is one of the most common causes for smoke, we can easily see that P(Fire | do(Smoke)) would probably be a lot lower (as long as we don't use fire to create the smoke in which case P would obviously be 1) as smoke is caused by fire but does not cause it itself. Therefore P(Fire | do(Smoke)) should be about the same as P(Fire).

When applying this to path diagrams, using the do-operator is the same as erasing all arrows that point towards the varibale that we perform the intervention on. Applied to the example of fire and smoke this means that, while we normally have a diagram like this: Fire -> Smoke where the arrow leading from Fire to smoke signifies a very high probability that a fire causes smoke, the do-operator erases this connection which is why they are suddenly independent. Applied to the example of the cholera epidemic, P(Cholera | Water Company, Water Purity) was what Snow was collecting data on. Now if one was to devise a RCT and intervene on water purity, one would no longer have to control for water company as one has basically taken their job away; this would mean P(Cholera | Water Company, do(Water Purity)) = P(Cholera | do(Water Purity)).

And this is effectively the first axiom of the Do-Calculus, which allows us to no longer control for variables that don't affect the outcome:

P(Y | do(X), Z) = P(Y | do(X))

In this case the variable Z does not have an arrow leading to Y and therefore no longer has to be controlled for.

Let's look at the example of the cholera epidemic again but in this case none of the confounders are unobservable or unmeasurable. In this case we can control for all of them and calculate the probability P(Cholera | do(Water Purity), Hygiene, Poverty, etc) and we can see that we no longer have to intervene as the afroementioned back-door criterion has been fulfilled and there are no more confounders left that are not being adjusted for.

This is the second axiom which says that if one controls for a set of variables that closes all backdoor paths into the variable that we intervene on, we no longer have to intervene on it and do(X) = see(X):

P(Y | do(X), Z) = P(Y | X, Z)

Here, controlling for Z would condition for all confounders.

Now let's assume we were an academic rival of John Snow and we wanted to prove our theory that cholera was caused by polluted air, also known as miasma, and we would create a RCT before they were invented. This means we would have two groups of test subjects, which live under completely identical conditions except that one group can breathe in very clean air and the other one only very polluted air. This way we could calculate P(Cholera | do(Miasma)). However since Miasma was not the reason for cholera infections, we would find out that both groups would have about the same amounts of cholera cases, even though one would probably have more cases of lung diseases. The result would therefore be P(Cholera | do(Miasma)) = P(Cholera).

This is essentially what the third axiom states. If we intervene on a variable that has no effect on the outcome, we could just remove that variable from the equation and get the same result:

P(Y | do(X)) = P(Y)

In this case X is our treatment variable that is actually not affecting the outcome Y.

With these three axioms we are generally able to transform any rung two question into a rung one question, meaning one should never be forced to rely on conducting a RCT, which even though often very pratical, can also be ethically or morally questionable or simply economically impossible to see through. The same however does not hold true for questions that belong on the third rung of the ladder of causation.

This level of causal thinking is concerned with the question of what would have been if things had gone differently, in short counterfactuals. This is the level that requires the most abstract thinking, and according to Pearl is what allowed humans to become such an intellectually advanced species, setting themselves apart from any other living beings.

This level of abstractivity however also requires a prize when it comes to the simplicity of finding the right answers. While first and second rung questions can be answered by simply performing an experiment, setting up an experiment to answer "What would have been if...?" question would require us to be able to turn back time; which, unfortunately(?), we aren't. This is one of the main problems that statistics historically had with counterfactuals and also probably the reason why it had such a bad relationship with causality. After all the "Why"-Question is very closely connected to counterfactuality. Asking for causes implies that nature adheres to certain rules which would allow the creation of a model and with that the simulation of alternative worlds where certain variables would have been different. This strongly contradicts the approach of the classic statistics of the late 19th and the 20th century of trying to keep statistics free of models and the idea that everything can be gained from only looking at the data.

At this point I can't help but draw a line to austrian Philosophist Karl Popper, in whose scientifical and philosphical tradition I see Pearl. Popper's field was not statistics but the philosophy of science, however those two fields definetely have quite a few connective points. In his famous book *The Logic of Scientific Discovery* Popper introduces the principle of falsification, under which in his opinion science should operate. According to this principle the task of science would be to postulate hypothesises and then do everything to try and disprove them and the value of a theory would be measured by how often and how strongly it was tried and not falsified. Obviously if a hypothesis was falsified it would have to be abandoned. This theory was a stark contrast to that of his academic opponents, the so called Positivists, who believed that a scientist's job was to observe and document nature and then accumulate his findings to postulate a theory which he should try to justify and prove using the principle of induction. For example a positivist would observe swans for a long time and only see white swans and could then use the principle of induction (e.g. the uniformity of nature) to prove that all swans are white. The weak point in this logic was quite obviously the principle of induction which lacked logical justifications and created contradictions and was the main point of Popper's attack on Positivism. Another argument of Popper against Positivism was the fact that implications, which are the forms rules of nature take, can only be falsified but never verified, since no matter how many examples one might find where a certain rule applies, one may never be certain that one won't find a counter example; and after all one counter example is enough to falsify an implication while it can never be verified, no matter how many positive examples one might find.
In this way Positivism can be compared to the classic statisticians as they want to only rely on data and shy away from making presumptions that they cannot justify with data as they are afraid of being accused of being non-empirical. They see metaphysics as useless and believe only purely empirical science is worth practicing (page 28 *The Logic of Scientific Discovery*), while Popper acknowledges the fact that he cannot gain true knowledge and that he is not interested in data itself but rather in trying to understand how the world works. To that he is happy to make presumptions that he expects will turn out to be false, but nonetheless by doing so achieving progress and learning more about the world. In this way Popper and Pearl share common traits as they both advocate a model based approach to science and both are of the opinion that one has to make strong assumptions in order to progress and should not be afraid of being wrong, but rather use it as a learning opportunity.

I would like to conclude by describing an example from Pearl's book which shows very nicely how

counterfactuals and causal models can play a big role in answering current pressing questions. "Is [insert weather catastrophe] caused by climate change?" is a question that one reads quite often these days, the answers however are often rather timid. On the one hand extreme weather events are obviously nothing new on the other hand they are trending upwards in both frequency and severity. More clarity can be achieved by using the counterfactual concepts of Probability of Necessity (short PN) and Probability of Sufficiency (short PS) which estimate how likely something was necessary or sufficient to cause an event. When using these terms in cases of extreme weather events, we suddenly get very clear and drastic answers. When one asks whether climate change was necessary to cause the 2003 European heatwave for example, the answer will be yes with a probability of 0.9, the likelihood of it being sufficient however is only 0.0072. These answers make a lot of sense if one looks at a causal model of what affects weather events. On one hand there is the natural variability which basically symbolises the fact that weather does not stay the same but instead is constantly changing, sometimes leading to extreme conditions and then there is the factor of climate change which for the most part increases the overall temperature. Now if one looks at the 2003 heatwave, it was likely that climate change was necessary for it to occur since internal weather variability was unlikely to cause such high temperatures on its own, but it couldn't be sufficient as that would mean internal variability wouldn't be necessary for the heat wave to occur in which case the temperatures of the heat wave would have been the standard temperatures, which obviously wasn't the case since then it wouldn't have been a heatwave anymore. Therefore we are clearly interested in the PN and the question of "Was climate change responsible for the 2003 heatwave?" can be answered with "Yes, it was, with a likelihood of 90%."
This demonstrates the importance of using a model-based approach when analysing data as without one, one could not create this conclusion but would instead only be able to recite how the data looks and that one cannot draw conclusions from it.

Sources

Pearl, Judea and Mackenzie, Dana. *The Book of Why* Penguin Books, 2019

Popper, Karl. *Logik der Forschung* Tübingen: Mohr Siebeck, 2005 (*Logic of Scientific Discovery)*

Internet Sources concerning Alphastar:

https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

https://www.youtube.com/watch?v=ZsCnuDgDcPo