Causal Data Science – notebook

Table of Contents

[Lecture 1: Intro to the course 3](#_Toc86153452)

[*Topic: causal knowledge for decision-making, ladder of causation* 3](#_Toc86153453)

[Introduction – Pearl, J., and D. Mackenzie (2018). The Book of Why. Basic Books, New York. 3](#_Toc86153454)

[Causal Inference in statistics, chapter 1: Statical & causal models 3](#_Toc86153455)

[Lecture 2: Graphical Causal models I 5](#_Toc86153456)

[*Topic: DAGs, D-seperation & testable implications, interventions in SCMs, backdoor criterion* 5](#_Toc86153457)

[Chapter 1 - The Book of Why, “the ladder of causation” 5](#_Toc86153458)

[Causal Inference and Data Fusion in Econometrics: section 1 & 2 5](#_Toc86153459)

[Causal Inference in statistics, chapter 2: Graphical models & their application 6](#_Toc86153460)

[Lecture 3: Software Exercises 8](#_Toc86153461)

[Chapter 2 - The Book of Why, “the genesis of causal inference” 8](#_Toc86153462)

[Lecture 4: Graphical Causal Models II 8](#_Toc86153463)

[*Topic: Matching & Regression, front-door criterion, causal discovery* 8](#_Toc86153464)

[Chapter 3 - The Book of Why, “from evidence to causes” (\*missing) 8](#_Toc86153465)

[Sections 3.1 & 3.2 Causal inference: confounding bias & covariate selection&backdoor 9](#_Toc86153466)

[Causal Inference in statistics, chapter 3: Graphical models & their application (\*missing) 10](#_Toc86153467)

[Pearl (1995) causal diagrams for empirical research (\*missing) 10](#_Toc86153468)

[Neal, the PC Algorithm for Causal Discovery 10](#_Toc86153469)

[Lecture 5: Experiments 11](#_Toc86153470)

[*Topic: RCTs, A/B Testing, difference-in-differences* 11](#_Toc86153471)

[The book of why, Chapter 4: Confounding and deconfounding: or, slaying the lurking variable (\*missing) 11](#_Toc86153472)

[The suprising power of online experiments [link] 11](#_Toc86153473)

[Avoid the pitfalls of A/B testing [link] 12](#_Toc86153474)

[Mixtape, Chapter 9: difference-in-differences (\*missing) 13](#_Toc86153475)

[Lecture 6: Surrogate experiments (z-identification) 13](#_Toc86153476)

[*Topic: Z-identification, instrumental variables, RDDs* 13](#_Toc86153477)

[The book of why, Chapter 5: The smoke-filled debate, clearing the air 13](#_Toc86153478)

[Does a long-term orientation create value? Evidence from a regression discontinuity 13](#_Toc86153479)

[Mixtape: Chapter 6 – Regression discontinuity (\*Missing) 13](#_Toc86153480)

[Mixtape: Chapter 7 – Instrumental variables (\*missing) 14](#_Toc86153481)

[P. Hü (2021) - Section 3.4: Identification by surrogate experiments (\*missing) 14](#_Toc86153482)

[Notes taken during the lecture 14](#_Toc86153483)

[Lecture 7: Application of Causal Inference in business 17](#_Toc86153484)

[In defense of curve-fitting: lyft keynote (rewatch) 17](#_Toc86153485)

[The book of why, Chapter 6: Paradoxes Galore (\*missing) 17](#_Toc86153486)

[Causal Machine Learning and Business decision making variable paper (\*missing) 17](#_Toc86153487)

[How to push causal inference in industry? paper (\*missing) 17](#_Toc86153488)

[Lecture 8: Causal artificial intelligence 17](#_Toc86153489)

[*Topic: Do-calculus, identification algorithms, data fusion paradigm* 17](#_Toc86153490)

[Causality: testing identifiability 17](#_Toc86153491)

[What is Causal Data Fusion? 18](#_Toc86153492)

[P. Hü (2021) section 3.3 – Causal calculus and the algorithmitization of identification strategies (\*missing) 18](#_Toc86153493)

[The book of why, Chapter 7: Beyond Adjustments, The conquest of mount interventions (\*missing) 19](#_Toc86153494)

[Lecture 9: Sample selection bias 19](#_Toc86153495)

[*Topic: Selection diagrams, recovering from selection bias in causal diagrams, selection propensity score, Heckman selection model* 19](#_Toc86153496)

[The book of why, Chapter 8: Counterfactuals – the world that could have been (\*missing) 19](#_Toc86153497)

[P. Hü (2021), section 4: Sample selection bias (\*missing) 19](#_Toc86153498)

[Lecture 10: Counterfactuals 19](#_Toc86153499)

[*Topic: Potential outcomes framework, ignorability, mediation & causal mechanisms, fairness in algorithmic decision-making* 19](#_Toc86153500)

[The book of why, Chapter 9: Mediation – the search for a mechanism (\*missing) 19](#_Toc86153501)

[The book of why, Chapter 6: Paradoxes Galore (\*missing) 20](#_Toc86153502)

[Lecture 11: External Validity 20](#_Toc86153503)

[*Topic: External Validity, the transportability problem, meta-transportabilty* 20](#_Toc86153504)

[The book of why, Chapter 10: Big Data, AI & The Big Questions (\*missing) 20](#_Toc86153505)

[Section 5 in Hünermund & Bareinboim (2021) (\*missing) 20](#_Toc86153506)

[Townsend, J. (Microsoft). A/B Testing and Covid-19: Data-Driven Decisions in Times of Uncertainty [link] (\*missing) 20](#_Toc86153507)

# Lecture 1: Intro to the course

## *Topic: causal knowledge for decision-making, ladder of causation*

## Introduction – Pearl, J., and D. Mackenzie (2018). The Book of Why. Basic Books, New York.

The main claim is that we deal to much with correlation and too little with causation, while correlation can produce some incredible results (NN, despite being all about fitting a function); it will not put us on the path to real machine intelligence. While the way that will is yet to be fully paved, he says:

* The calculus of causation consists of two languages: causal diagrams, to express what we know, and a symbolic language, resembling algebra, to express what we want to know

**Do-operator**

* P(L|Do(x)) or P(L|do(not-x)): the effect of drug (x) on lifespan (L) if made to take the drug, or opposite, not to the take drug.
  + This is known as an *intervention or treatment*, and the do-operator signifies dealing with an intervention rather than a passive observation.

**The example: causal inference engine**

* Accepts three different kinds of inputs: assumptions, queries, and data: producing 3 kind of outputs:

1. Yes/No, whether the given query in theory can be answered under the existing causal model, a summing perfect and unlimited data
2. If yes, it produces and Estimand – a formula that is the recipe for generating the answer from any hypothetical data, when available
3. Given the input and the estimand, it produces an actual estimate for the answer, along statistical estimates indication the amount of uncertainty.

## Causal Inference in statistics, chapter 1: Statical & causal models

**Simpson’s paradox:** existence of data in which a statistical association that holds for an entire population is reversed in every subpopulation, e.g. among the drug-takers a lower percentage recovered, but if we partition by gender, we see more men taking the drug recover than men not taking the drug, and more women taking the drug recover than do women not taking it. The reason being, if one gender is more likely to take the drug and/or less likely to recover.

**First working definition of causation: *“****variable X is a cause of a variable Y if Y in any way relies on X for its value”*

**Second working definition of causation:** “*A variable X is a direct cause of a variable Y if X appears in the function that assigns Y’s value. X is a cause of Y if it is a direct cause of Y, or of any cause of Y.”*

**Structural Causal Models (SCM)**

* Two sets of variables, U and V, and a set of functions f assigning each variable in V a value based on the value of the other variables in the model
* Variables in U are called *exogenous* variables, i.e., external to the model. And for whatever reason, we choose not to explain how they are caused. (Cannot be descendants, particularly not of V)
* Variables in V are called *endogenous,* every of such variable is a descendant of at least one U (represented as root nodes and have no ancestors)
* Directed acyclic graphs (DAGs), to see if a🡪b then b depends on a, or formally for the causation in graphs: *“If, in a graphical model, a variable X is the child of another variable Y, then Y is a direct cause of X; if X is a descendant of Y, then Y is a potential cause of X. There are rare intransitive cases in which Y will not be a cause of X”*

**Rule of product decomposition**

*“For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions P(child|parents) over all the “families” in the graph.”*

# Lecture 2: Graphical Causal models I

## *Topic: DAGs, D-seperation & testable implications, interventions in SCMs, backdoor criterion*

## Chapter 1 - The Book of Why, “the ladder of causation”

**The three levels of causation (p.34)**

1. **Association**

**Activity:** seeing, observing

**Questions:** what if I see? Would seeing X change my belief in Y?

**Examples:** What does a symptom tell me about my disease? What does a survey tell us about the election result?

X

1. **Intervention**

**Activity:** Doing, intervention

**Questions:** What if I do? How? What could Y be if I do X? How can I make Y happen?

**Examples:** If I take the aspirin, will my headache be cured? What if we ban cigarettes?

X

1. **Counterfactuals**

**Activity:** Imagining, Retrospection, Understanding

**Questions:** what if I had done? Why?

**Examples:** was it the aspirin that stopped my headache? Would JFK be alive if LHO had not killed him

X

## Causal Inference and Data Fusion in Econometrics: section 1 & 2

One key feature of DAGs is that they are falsiable through testable implications over the observed distributions, including conditional independence relationships between variables in the model

* D-separation allows to systematically read of the conditional independencies implied by the structural model from the graph.
  + This method provides the analyst with a set of testable implications that can be benchmarked with the available data.
* These independence relations can easily be tested through statistical hypothesis testing, and if rejected, the hypothesized model can be discarded too.
* An advantage of such local tests, compared to global goodness-of- t measures, for example, that they indicate exactly where the model is incompatible with the observed data. Thus, the analyst can rely on concrete clues about where to improve the model, which facilitates an iterative process of model building.
* **As such, the causal model is using d-seperation to the determine the conditional independencies and uses that to elude a set of testable implications. These can be tested via statistical hypothesis testing, and if rejected, so is the hypothesized model. Compared to global tests, such as t measures, this has the benefit of indicating exactly where the model is incompatible since the test is done locally and thereby allowing for iterative model building.**

*Interventions in SCM*

* *“The aim of causal inference is to predict the effects of interventions, such as those resulting from policy actions, social programs, and management initiatives”*
  + Interventions are done by deleting individual functions and/or fixing the left-hand side variables at a constant value (do-operator)

## Causal Inference in statistics, chapter 2: Graphical models & their application

*“The definition of conditioning as filtering by the value of the conditioning variable.*

*When we condition on Z, we limit our comparisons to cases in which Z takes the same value.*

*But remember that Z depends, for its value, on X and Y.”*

* + Conditioning on a variable along a chain or fork blocks (\d-separates") the path
* Conditioning on a collider opens the path
* IF two variables are conditionally independent then there correlation coefficient is 0

**Rule 1 (conditional independence in chains):**

* *Two variables, x and y, are conditionally independent given z, if there is only one unidirectional path between x and y, and z is any set of variables that intercepts that path*

**Rule 2 (conditional independence in forks):**

* *If a variable x is a common cause of variables y and z, and there is only one path between y and z, then y and z are independent conditional on x*

**Rule 3 (colliders):**

* Each cause and the outcome is likely dependent
* The respective causes are likely independent
* The causes are likely dependent conditioned on the outcome
  + Conditioned on Z, if Z=10 and X=3, then given X and Y are common causes of Z, Y must equal = 7
  + The same holds for descendants of the collider
  + *conditioning on a collision node produces a dependence between the node’s parents*

**D-separation**

* A criterion or process that is applicable to a graphical causal model of any complexity in order to predict dependencies shared by all data sets generated by that graph
* “D” stands for directional, and determines for any pair of nodes, whether a connecting path exist (d-connected) or not (d-separated)
  + D-separated means the variables are definitely independent
  + D-connected means the variables are possibly, or most likely, connected
* Fully blocked paths equals d-separation, but whether blocked depends on it being unconditioned or conditioned
  + If not conditioning, only colliders can block a path
  + If conditioning, two potential blockers:
    - colliders conditioned on Z (not in Z), and has no descendants of Z
    - A chain or fork whose middle node is in Z

Text, letter

Description automatically generated

**Model testing and Causal Search**

* The testable implications derived of any causal model

Diagram

Description automatically generated

We can demonstrate it on the causal model of Figure 2.9. Among the many conditional independencies advertised by the model, we find that W and Z1 are independent given X, because X d-separates W from Z1. Now suppose we regress W on X and Z1. Namely, we find the line

w = rXx + r1z1

that best fits our data. If it turns out that r1 is not equal to zero, we know that W depends on Z1 given X and, consequently, that the model is wrong. [Recall, conditional correlation implies conditional dependence.] Not only do we know that the model is wrong, but we also know where it is wrong; the true model must have a path between W and Z1 that is not d-separated by X. Finally, this is a theoretical result that holds for all acyclic models with independent errors (Verma and Pearl 1990), and we also know that if every d-separation condition in the model matches a conditional independence in the data, then no further test can refute the model. This means that, for any data set whatsoever, one can always find a set of functions F for the model.

# Lecture 3: Software Exercises

## Chapter 2 - The Book of Why, “the genesis of causal inference”

White developed the realm based upon a study of heredity of guniea pigs. Fisher, Niles, K/Carl Pearson all great opponents.

# Lecture 4: Graphical Causal Models II

## *Topic: Matching & Regression, front-door criterion, causal discovery*

## Chapter 3 - The Book of Why, “from evidence to causes” (\*missing)

www

## Sections 3.1 & 3.2 Causal inference: confounding bias & covariate selection&backdoor

**Confounding bias**

* One of the biggest threats to causal inference: a correlation might not reflect a genuine causal link between two variables
* In the presence of confounding, a non-trivial solution must be found through the use of do-calculus, two special cases for dealing with this are instances of the general treatment provided by do-calculus; backdoor and frontdoor criterion

**Covariate selection & backdoor criterion**

* A set Z is backdoor admissible if it blocks every path between X and Y in the graph Gx
* Identification via backdoor adjustment requires that all backdoor paths can be blocked by a set of observed nodes, which is not always feasible in many practical settings
* Unblocked paths between X and Y pointing into X (i.e., \entering through the backdoor") create an association between X and Y that is not due to any causal inuence exerted by X.17
* By adjusting for (or conditioning on) variables along these spurious paths, this association can be canceled such that only the causal inuence from X to Y remains

*Theorem:* if a set of variables satisfies the backdoor criterion relative to (x,y), the causal effect of x on y can be identified. Practically speaking, estimation can be carried out by propensity score matching, inverse probability weighting, deep neural networks, or weighted empirical risk minimization, among other efficient estimation methods.

**Front-door adjustments in the presence of unmeasured confounders**

* If unobserved confoudners remains in the graph, no backdoor adjustment set can do the job, front-door criterion comes to play:
* A set of variables Z is said to satisfy the conditional front-door criterion relative to a triplet (X,Y,W) if
  + Z intercepts all directed paths from X to Y
  + No unblocked backdoor path from X to Z given W
  + All the backdoor paths from Z to Y are blocked by {X,W}
* Then, the causal effect of X on Y can be identified

## Causal Inference in statistics, chapter 3: Graphical models & their application (\*missing)

xxx

## Pearl (1995) causal diagrams for empirical research (\*missing)

Xxx

## Neal, the PC Algorithm for Causal Discovery

A method to discover the graph from conditional independencies, or more often, identify the markov equiveillance class.

Step 1 – identify skeleton

* Start with undirected graphs and remove X – Y where X is conditionally independent of Y given some conditioning set Z (even null set), starting from null and increasing

Step 2 – identify immoralities

Text

Description automatically generated

Step 3-

A picture containing diagram

Description automatically generated

# Lecture 5: Experiments

## *Topic: RCTs, A/B Testing, difference-in-differences*

## The book of why, Chapter 4: Confounding and deconfounding: or, slaying the lurking variable (\*missing)

Confounding bias, when Daniel did not consider him and his friends current health compared to the control group

While confounders are the issue in deriving causality, deconfounders are the cure

* 1) confounding needs and has a causal solution, and (2) causal diagrams provide a complete and systematic way of finding that solution
* Knowing the effects of the confounding variable, makes it easy to “adjust for Z” / “control for Z”

**Randomization**

Thus, randomization actually brings two benefits. First, it eliminates confounder bias (it asks Nature the right question). Second, it enables the researcher to quantify his uncertainty

* RCTs main purpose is to eliminate confounding

**The new paradigm of confounding**

MISSING THE LAST 5-10 PAGES OF THE CHAPTER

## The suprising power of online experiments [[link](https://hbr.org/2017/09/the-surprising-power-of-online-experiments)]

Three ways to organize the company’s experimentation model:

1. **Centralized**, a team of data scientists serve the entire company
2. **Decentralized,**
3. **Center-of-excellence model**

* Needs to establish *overall evaluation criterion* to measure the experiment by
* Data quality is of the outmost importance, e.g. run A/A test to confirm no change in 95% of the cases

*“If you really want to understand the value of an experiment, look at the difference between its expected outcome and its actual result. If you thought something was going to happen and it happened, then you haven’t learned much. If you thought something was going to happen and it didn’t, then you’ve learned something important. And if you thought something minor was going to happen, and the results are a major surprise and lead to a breakthrough, you’ve learned something highly valuable.”*

## Avoid the pitfalls of A/B testing [[link](https://hbr.org/2020/03/avoid-the-pitfalls-of-a-b-testing)]

**Summary:** “*Often firms make serious mistakes in conducting these tests: They focus on the average, instead of looking at how a change impacts different customer segments. They forget that customers are connected and that their interactions can affect test outcomes. And they run tests for too short a period, failing to recognize that customers’ reactions can change over time.”*

**Pitfall 1: Not looking beyond average**

* Considering a fictional average user, but one segment may greatly improve, while another greatly decrease, hence canceling each other out on average

Essentially, it comes down to the heterogeneity of the different users, which may require different solutions for different segments, i.e., firms need to:

1. Use metrics and approaches that reflect the value of different segments
   1. Netflix uses interleaving A/B designs, where the user’s experience is alternated between A and B,
2. Measure the impact across different levels of digital access
   1. Segregating with respect to the user’s digital environment, e.g. internet connection, device etc.
3. Always account for group-specific behaviour
4. Segment key markets
   1. In India, much of the online access comes from mobile devices and networks, which means any heavy feature is less values in India than in the US

**Pitfall 2: Forgetting that customers are connected**

* Use network based A/B
* Use time-series experiments

**Pitfall 3:** Focusing predominantly on the short-term

* Need to get the duration of experiments right, so novelty effects and steady-steady stage are measured for any new feature

## Mixtape, Chapter 9: difference-in-differences (\*missing)

XXX

# Lecture 6: Surrogate experiments (z-identification)

## *Topic: Z-identification, instrumental variables, RDDs*

## The book of why, Chapter 5: The smoke-filled debate, clearing the air

Since it was not until the 90’s we got a language for causality, it was immensely difficult during the 60-90’s to prove smoking caused lung cancer. Interesting approaches:

* RCT, not possible nor ethical to have a certain group smoke for 10 years, in case the hypothesis proved right
* Crazy large surveys
* Control & Treatment group for an interview to determine any diagnoses of cancer and habits of smoking

Fisher interestingly argued that it was impossible to measure, as an underlying difference in genome might be the reason, and hence those with this genome were more inclined to smoke but irrespective of the smoke, the disease would have gotten them anyways. Mathematically this was disputed by taking the stance: “If such genome exists, how much can it, and not the smoking, explain of the lung cancer cases?”

## Does a long-term orientation create value? Evidence from a regression discontinuity

Conducts causal study using the RDD methodology, whilst highlighting the tradeoff internal versus external validity, they find long-term orientation is value-enhancing for the company. On the day of passing long-term proposals; abnormal returns are detected. More sustainably, this combats myopia in management by better aligning goals and compensation.

## Mixtape: Chapter 6 – Regression discontinuity (\*Missing)

While a method from 1960 is only now being used frequently, because of its ability to convincingly eliminate selection bias

## Mixtape: Chapter 7 – Instrumental variables (\*missing)

Dssds

## P. Hü (2021) - Section 3.4: Identification by surrogate experiments (\*missing)

Dssds

## Notes taken during the lecture

A diagram of a compass

Description automatically generated with low confidence

* why is 🡪 Not backdoor-admissible
* Because one of these paths will always remain open
* Chain is open, collider is closed
* An experiment or intervention (to X) remove all the edges pointing to that point (x), whether directional or bidirectional *[the connected and removed edges can be drawn as pre/post intervention]*
* The purpose is to derive the above equation, by experimentation on Z

**Identification by surrogate experiment**

* Three variables, disjoint sets,

A picture containing text

Description automatically generated

* Q identifiable in G: if that’s the case I should be able to find a backdoor-criterion before hand
* (ii) Find a note such that,
  + (a) x (treatment variable) intercepts all direct effect between two variables
  + (b) Q is identifiable in 🡪 can the effect between the Z and Y be identified?

**Same point, where does the definition hold:**

**Diagram

Description automatically generated**

1. This is the one, since the effect can be isolated to Z 🡪 X 🡪 Y (a chain, hence no backdoor path)
2. Experimenting on Z, would still leave bi-directional unobserved confounder for x <-> y and thus violating condition B from above
3. Since X does not intercept all the paths and thus violating condition A from above

**Instrumental variables**

* When Z-identification is not possible
* Id estimator
* By converting from non-parametric to linearity (somehow)

**Local average treatment effect**

* X: military service (yes / no)
* Z: drafted (yes / no)
  + 4 Groups: compliers (doing as told), defies (doing the opposite), always takers, never-takers
* By ID estimation we can only say something about the complier group (ruled out defiers)
* However, the problem is: the latent subgroup (you can say which is which ahead of them being drafted)
* “Weak instrument problem” 🡪 statiscal problem

**Regression discontinuity design**

* A form of ID estimation
* To diminish the difference between companies and thereby confounding variables causing this difference, take a group of 49% to 51% of the votes, that is very close neighborhood to the cutoffs = not confounded.

*Two designs of RDD*

1. Sharp
   1. Probability to receive treatment jumps from zero to one at the discountinity
2. Fuzzy
   1. Not neceasrily a zero percent treatment possibility at the boundary, but changes drastically

*Estimation*

* Restrict to the neighborhood, Identify the threshold and show that there is effect
  + Cost: the smaller the neighborhood, the less bias, but the more variance (due to the less data): variance-bias trade-off 🡪 The art of RDD is to balance this trade-off
  + Issues: “bunching” below or above the threshold (able to manipulate the running variable e.g. persuade teachers to give minimum passing grades)
    - Plot the data to check for bunching in the form of discountinuity

# Lecture 7: Application of Causal Inference in business

## In defense of curve-fitting: [lyft keynote](https://causalscience.netlify.app/programme/keynote-videos) (rewatch)

## The book of why, Chapter 6: Paradoxes Galore (\*missing)

XX

## Causal Machine Learning and Business decision making variable [paper](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867326) (\*missing)

Confounding bias, when Daniel did not consider him and his friends current health compared to the control group

## How to push causal inference in industry? [paper](https://causalscience.org/blog/how-to-push-causal-inference-in-industry) (\*missing)

Confounding bias, when Daniel did not consider him and his friends current health compared to the control group

# Lecture 8: Causal artificial intelligence

## *Topic: Do-calculus, identification algorithms, data fusion paradigm*

## [Causality: testing identifiability](https://david-salazar.github.io/2020/07/31/causality-testing-identifiability/)

Often, front-door criterion or backdoor criterion will identify causal models, but not always due lack of necessary variables. When identifiable, the model is said to be markovian: any causal effect X 🡪 Y is identifiable if we have measurements of the parents of X, Pa(X)

“We must remember that our causal model is simultaneously a probabilistic model. In particular, they induce a **decomposition of the joint probability distribution** because each variable is independent of all its non-descendants given its direct parents in the graph. However, when our model contains *unobserved confounders*, we must *marginalize them* in order to obtain the joint probability distribution of the observed variables”

”What is the intuition of our identifiability test? The key to identifiability *lies not* in blocking back-door paths between X and Y but, rather, in **blocking back-door paths between**XX**and any of its descendants that is an ancestor of**YY. Thus, by blocking these paths, we can ascertain *which part of the association* we observe is **spurious** and which *genuinely causative*.”

## [What is Causal Data Fusion?](https://www.causalscience.org/blog/what-is-causal-data-fusion/)

The automated and do-calculus based inferential engine that opens up a complete causal inference pipeline from problem specification, to identification, to estimation

It consists of three syntactic rules that allow us to manipulate causal queries and transform them into other observable expressions, identification algorithms take three inputs

* A target query
* A causal model that captures our domain knowledge of the context under study
* And the type of data available to us

Then, whenever a solution exist, such is given, and one can move to estimation of the desired causal effect, e.g. via weighted empirical risk minimization

* [In-depth paper of this](https://arxiv.org/abs/1912.09104)

## P. Hü (2021) section 3.3 – Causal calculus and the algorithmitization of identification strategies

The backdoor and front-door only represent a subset of the overall identification results derivable in DAGs. In more generality, identifiable of any query in the form can be decided systematically from causal inference engine called do-calculus.

**Do-Calculus**

Do-calculus consists of three inference rules, allowing to transform probabilistic sentences involving interventions and observations, whenever certain separation conditions hold in the causal graph G defined by the model M

*Notation:*

Let X, Y , Z, and W be arbitrary disjoint sets of nodes in G.

* The mutilated graph that is obtained by removing all arrows pointing to nodes in X from G is denoted by (X received no causal effects in the DAG).
* Similarly, results from deleting all arrows that are emitted by X in G (X sends no causal effects in the DAG).
* Finally, the removal of both arrows incoming in X and arrows outgoing from Z is denoted by (X neither sends nor received causal effects in the DAG).

Given this notation, the following three rules – valid for every interventional distribution compatible with G – can be formulated:

*Rules:*

Text, letter

Description automatically generated

Rule 1: a reaffirmation of the d-seperation criterion for the X-manipulated graph . Since Z is independent of Y, conditional in X and W, Z can be freely inserted or deleted in the do-expression

Rule 2: states the condition for an intervention do(Z = z) to have same effect as a passive observation Z = z. The condition is fulfilled if blocks all backdoor paths from Z to Y. Note that only such backdoor paths are remaining here, since edges emitted by Z are deleted from the graph.

Rule 3: indicates under which condition a manipulation of Z does not affect the probability of Y. This is the case if in the X- and Z-manipulated graph, Z is independent of Y conditional on X and W

Identifiability of a causal query can be decided by repeatedly applying the rules of do-calculus, until Q is transformed into a final expression that no longer contains a do-operator. This provides the basis for consistent estimation of Q from nonexperimental data.

Example use of do-calculus:

Diagram

Description automatically generated

## The book of why, Chapter 7: Beyond Adjustments, The conquest of mount interventions

**To climb the mountain of counterfactuals, the most novice climber may use back door criterion or its cousins, wheras do-calculus works in any instance:**

For many researchers, the most (perhaps only) familiar method of predicting the effect of an intervention is to “control” for confounders using the. adjustment formula. This is **the method to use if you are confident that you have data on a sufficient set of variables (called deconfounders) to block all the back-door paths between the intervention and the outcome**.

* To do this, wemeasure the average causal effect of an intervention by first estimating its effect at each “level,” or stratum, of the deconfounder.
* We then compute a weighted average of those strata, where each stratum is weighted according to its prevalence in the population. If, for example, the deconfounder is gender, we first estimate the causal effect for males and females.
* Then we average the two, if the population is (as usual) half male and half female. If the proportions are different—say, two-thirds male and one-third female—then to estimate the average causal effect we would take a correspondingly weighted average.

**Regression & back-door**

If we have confounders then fitting a regression line will only give us the average observed trend, and not the average observed delta of Y given X. However, we can fit the best regression plane and nature will do all the averaging for us. That is,

The coefficient a in the equation of that plane, Y = aX + bZ + c, will automatically adjust the observed trend of Y on X to account for the confounder Z. If Z is the only confounder, then a is the average causal effect of X on Y. A truly miraculous simplification!

In short, sometimes a regression coefficient represents a causal effect, and sometimes it does not—and you can’t rely on the data alone to tell you the difference (rely then on the DAG)

*\*regression-based adjustment works only for linear models, in non-parametric cases more extrapolation is needed*

To sum up, the back-door adjustment formula and the back-door criterion are like the front and back of a coin. The back-door criterion tells us which sets of variables we can use to deconfound our data. The adjustment formula actually does the deconfounding. In the simplest case of linear regression, partial regression coefficients perform the back-door adjustment implicitly. In the nonparametric case, we must do the adjustment explicitly, either using the back-door adjustment formula directly on the data or on some extrapolated version of it. **However, if a backdoor path cannot be blocked, this will not work (frontdoor criterion might).**

**Front-door criterion**

It differs from the back-door adjustment in that we adjust for two variables (Smoking and Tar) instead of one, and these variables lie on the front-door path from Smoking to Cancer rather than the back-door path.

**Do-calculus**

In both the front- and back-door adjustment formulas, the ultimate goal is to calculate the effect of an intervention, P(Y | do(X)), in terms of data such as P(Y | X, A, B, Z,…) that do not involve a do-operator. If we are completely successful at eliminating the do’s, then we can use observational data to estimate the causal effect, allowing us to leap from rung one to rung two of the Ladder of Causation

## Notes from lecture

*The do-calculus rules*

1. Insert or delete of observations
   1. Y and Z needs to be d-seperated/independent given X, W – then we can get rid of z altogether
2. Action/observation exchange, replace do(z) with normally conditioned upon z without do(), if certain d-separation relationships hold
3. Insertion/deletion, we get rid of the do() all together

G-over-bar = delete edges pointing into X

G-under-bar = delete edges emitted by X

G-and-a-bar-under-the-note = delete edges emitted by X and pointing into X

**Exercise: do-calculus**

Chart

Description automatically generated

Target query:

1. Law of total prob:
2. Rule 2 of do-calculus, this holds in deleting edges emitted by and pointing to B

*Actual solution*

Target query:

* Identify two backdoor paths, A 🡪 B 🡪 C & D🡪B🡪C
  + A and D must be conditioned on since they are backdoor admissible paths
  + Answer

Do-calc

* In G-over-score-B, we obtain d-seperation, and thus it holds that A & D are independent of B.
* Then by rule 2, in G-under-score-B, it holds that C and B are d-seperated, given A,D

# Lecture 9: Sample selection bias

## *Topic: Selection diagrams, recovering from selection bias in causal diagrams, selection propensity score, Heckman selection model*

## The book of why, Chapter 8: Counterfactuals – the world that could have been (\*missing)

XX

## P. Hü (2021), section 4: Sample selection bias (\*missing)

XX

Knox, D., W. Lowe, and J. Mummolo (2020). Administrative Records Mask Racially Biased Policing (\*missing)

Sds

Conditional independence in sample selection models. Economics Letters (\*missing)

Xxx

Heckman, J. (1979). Sample Selection Bias as a Specification Error (\*missing)

Xx

# Lecture 10: Counterfactuals

## *Topic: Potential outcomes framework, ignorability, mediation & causal mechanisms, fairness in algorithmic decision-making*

## The book of why, Chapter 9: Mediation – the search for a mechanism (\*missing)

In summary, over the last fifteen years, the Causal Revolution has uncovered clear and simple rules for quantifying how much of a given effect is direct and how much is indirect. It has transformed mediation from a poorly understood concept with doubtful legitimacy into a popular and widely

applicable tool for scientific analysis.

Example

Fruit 🡪 Vitamin C 🡪 scurvy (if mediator, vitamin C, is unknown, that is if the mediator or underlying why is unknown, we can't fully understand what causes what)

The missing phrase in both their vocabularies was “hold constant.” Todisable the indirect path from Gender to Outcome, we must hold constant the variable Department and then tweak the variable Gender. When we hold the department constant, we prevent (figuratively speaking) the applicants from choosing which department to apply to. Because statisticians do not have a word for this concept, they do something superficially similar: they condition on Department. That was exactly what Bickel had done: he looked at the data department-by-department and concluded that there was no evidence of discrimination against women. That procedure is valid when Department and Outcome are unconfounded; in that case, seeing is the same as doing. But Kruskal correctly asked, “What if there is a confounder, State of Residence?” He probably didn’t realize that he was following in the footsteps of Burks, who had drawn essentially the same diagram. I cannot stress enough how often this blunder has been repeated over the years—conditioning on the mediator instead of holding the mediator constant. For that reason I call it the Mediation Fallacy. Admittedly, the blunder is harmless if there is no confounding of the mediator and the outcome. However, if there is confounding, it can completely reverse the analysis, as Kruskal’s numerical example showed. It can lead the investigator to conclude there is no discrimination when in fact there is.

**The direct effect**

Text

Description automatically generated

Text

Description automatically generated

**The indirect effect**

Text

Description automatically generated

**Mediation in Linear Wonderland**

* For linear causal model, counterfactuals are not neither, hence.

A picture containing text

Description automatically generated

In general, if there is more than one indirect pathway from X to Y, we evaluate the indirect effect along each pathway by taking the product of all the path coefficients along that pathway. Then we get the total indirect effect by adding up all the indirect causal pathways. Finally, the total effect of X on Y is the sum of the direct and indirect effects. This “sum of products” rule has been used since Sewall Wright invented path analysis, and, formally speaking, it indeed follows from the do-operator definition of total effect.

Satell, G. and Y. Abdel-Magied (2020).AI Fairness Isn’t Just an Ethical Issue [[link](https://hbr.org/2020/10/ai-fairness-isnt-just-an-ethical-issue)]

Very well written paper about the dangers of bias in AI, highlighted via prejudice in grades during covid, amazon’s recruitment algorithm, Steve Wozniak’s wife who had 10x worse credit score due to bias against women and the like.

Ethical paper as it claims 1) AI systems are tools designed to create value, not cause harm, for humans, 2) as with any human making decision impacting people’s life, oversight is needed, 3) it can in fact be profitable for a business to be fair.

Holland, P. (1986). Statistics and Causal Inference (\*Missing)

Xxx

Chapter 4 –Pearl, J., M. Glymour, and N. Jewell (2016). Causal Inference in Statistics –A Primer (\*Missing)

Xx

Zhang, J. and E. Bareinboim (2017). Fairness in Decision-Making –The Causal Explanation Formula [[link](https://causalai.net/r30.pdf)] (\*Missing)

Xxx

# Lecture 11: External Validity

## *Topic: External Validity, the transportability problem, meta-transportabilty*

## The book of why, Chapter 10: Big Data, AI & The Big Questions (\*missing)

XX

## Section 5 in Hünermund & Bareinboim (2021) (\*missing)

xxxx

## Townsend, J. (Microsoft). A/B Testing and Covid-19: Data-Driven Decisions in Times of Uncertainty [[link](https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/a-b-testing-and-covid-19-data-driven-decisions-in-times-of-uncertainty/)] (\*missing)

**From Extra video in Fusion**

*Z-identifiability*

* Need to tell fusion we have experimental data (surrogate)
* This makes it computable

*Selection bias*

* Add the selection node
* Whatever found in the sample is representative for the population as a whole, meaning the selection bias is not influencing

*Transportability*

**Pdfs**

“You can show, however, that you will never be able to perfectly determine a DAG

from data alone. Some ex-ante causal assumption will always be needed”

Slide 95 for discussion of the broader organizational perspective