Introduction to Causal Inference and Causal Data Science

Day 1: Causal Inference and Potential Outcomes

Oisín Ryan

Department of Data Science and Biostatistics
Julius Center
UMC Utrecht

July 1, 2025

Oisín Ryan

Wouter van Amsterdam

Bas Penning de Vries

Who are we

Oisín Ryan

Assistant Professor UMCU

- Causal Inference with Real World Data
  - Large-scale, multi-center, population health databases
- Co-ordinator causal inference, statistics and data engineering, RWE
- Applied research projects: Safety and Efficacy of Vaccines for regulatory bodies
- Previous background in social sciences
  - Causal effects of language ability on study success
  - Causal impact of after-school training programs in Rotterdam
  - Clinical Psychology applications

Who are we

Assistant Professor UMCU

- Intersection of Machine Learning and Causal Inference in Healthcare
- Using prediction models for decision making
- Individual treatment effect estimation in cancer settings
- Background in Physics (Bsc) Medicine (MD), Machine learning in Healthcare (PhD)

Wouter van Amsterdam

Who are we

Assistant Professor UMCU

- Methodological challenges in causal inference
- Causal Estimands in different study designs (e.g. case-control)
- Comparing methods for estimating time-varying effects
- Missing data, confounder adjustment and measurement error
- Currently focused on teaching statistics, methodology and causal inference

Bas Penning de Vries

This Course

Goals

- **Introduce** you to the foundational concepts of causal modeling with **observational** or **non-experimental** data.
- Show you how to **view data problems** or through lens(es) of causal inference
- **Equip you** with the **skills** to perform causal inference and causal modeling
- **Get you started** on your own journey of causal learning

This Course

Goals

- **Introduce** you to the foundational concepts of causal modeling with **observational** or **non-experimental** data.
- Show you how to **view data problems** or through lens(es) of causal inference
- **Equip you** with the **skills** to perform causal inference and causal modeling
- **Get you started** on your own journey of causal learning

Approach:

- **Broad** and **interdisciplinary**
- **Practical** and **hands-on**

Outline

1. Potential Outcomes
   - Adjusting for confounders I
2. Directed Acyclic Graphs
   - Adjusting for confounders II
3. Target Trial Emulation
   - Adjusting for confounders III
4. Causal Data Science
   - What is the difference between learning a causal and a predictive model from data?
   - How do ideas from causal modeling interact with prediction tasks?
5. Wrapping up
   - Longitudinal settings
   - Apply what you learned
   - Q & A

Practical Matters

Course materials and schedule:

- `https://tinyurl.com/y3z9c48p`
- Morning and Afternoon sessions: Lecture x 2, Practical x 1
- Practical are in **R**
- Early end on **Friday** (14.00)

Lunch is provided

- 12:00 - 13:00, brought here
- The room will not be locked during breaks: Take or leave possessions at your own risk

SCAN ME

Today

1. Why Causal Modeling?
   - How are statistical and causal modeling related?
   - What types of questions is causal modeling concerned with?
   - The "magic" of RCTs
2. Potential Outcomes
   - Framework for causal modeling
   - Individual and Average Treatment Effects
   - Assumptions for identifying causal effects
   - Identification vs Estimation
3. Adjustment Methods I: Stratification and Matching

Why Causal Modeling?

Why Causal Modeling?

- Does smoking cause cancer?
- Was the covid-vaccine effective in lowering the risk of covid?
- Is a medicine safe for pregnant women or does it lead to adverse health conditions?
- Does the expression of gene X produce phenotype Y?
- What is the effect of social media use on adolescent well-being?
- What effect could we expect a sugar tax to have on rates of adult-onset diabetes in the general population?
- What was the effect of covid-19 lockdowns on hospitalization numbers?
- Which treatment type will be most effective in reducing symptoms for this type of individual?

Why Causal Modeling?

- Does smoking cause cancer?
- Was the covid-vaccine effective in lowering the risk of covid?
- Is a medicine safe for pregnant women or does it lead to adverse health conditions?
- Does the expression of gene X produce phenotype Y?
- What is the effect of social media use on adolescent well-being?
- What effect could we expect a sugar tax to have on rates of adult-onset diabetes in the general population?
- What was the effect of covid-19 lockdowns on hospitalization numbers?
- Which treatment type will be most effective in reducing symptoms for this type of individual?

Causal Modeling: When can we answer causal questions using *data*? And how should we go about doing this?

Statistical vs Causal Modeling

Statistical modeling and **data science** give us a rich language to describe uncertainty in the world we see around us

- The language of *co-occurences*, *expected values*, *(joint, marginal and conditional) probabilities*, *statistical dependencies*, *predicted values*
- It helps us *describe* patterns and make (certain types of) *predictions*.

We will be using many ideas and tools from statistical modeling in this course

Example

Imagine we are a team of health scientists.

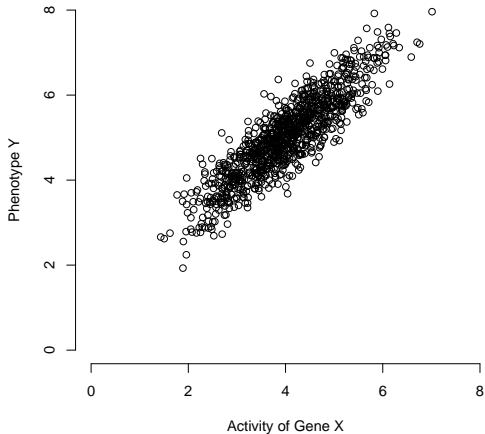We take a blood sample from a random sample of the population and record:

- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).

Example

Imagine we are a team of health scientists.

We take a blood sample from a random sample of the population and record:

- The level of expression of a particular gene X

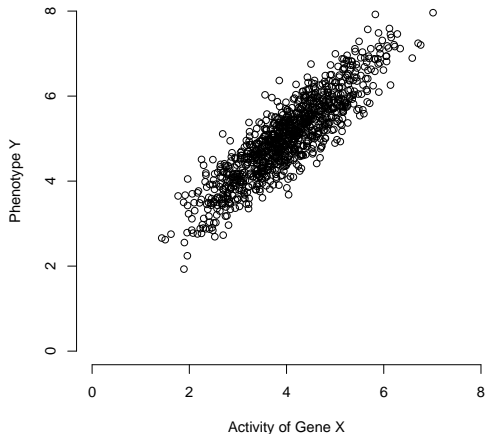- The level of expression of a phenotype Y (e.g. blood insulin levels).

Example

Imagine we are a team of health scientists.

We take a blood sample from a random sample of the population and record:

- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).
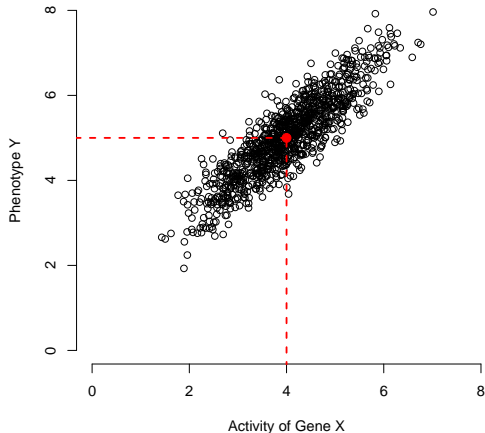

What kind of information can we extract from this data? What tasks can we perform, and what research questions can we answer using statistical techniques?
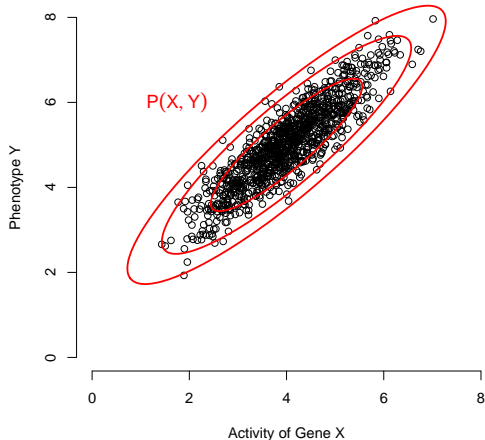
Example

Description:

- What is the average level of gene expression in the sample ($\bar{X}$)?

- What does that tell us about the average level in our population ($E[X]$)?

- How *certain* are we about our *estimate* of the population mean?

Example

Models for (co-)occurence

- What is the likelihood of observing a low level of gene expression (*Marginal Distribution* $P(X = x)$)?

- What is the probability that someone in the population has both a high insulin level and a high gene expression (*Joint distribution* $P(X, Y)$)?

- We can fit models, such as the normal distribution $P(X, Y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and ask how well this model fits the data

Example

Prediction:

- If I collect one more data point in *identical* circumstances and I observe a gene expression score of 5, what is my best guess of what phenotype level that person has?

- Answered by estimating / fitting models for the *conditional distribution* $P(Y|X)$

- Best guess is the *conditional expectation* $E[Y|X = 5]$, which we have to *estimate* somehow
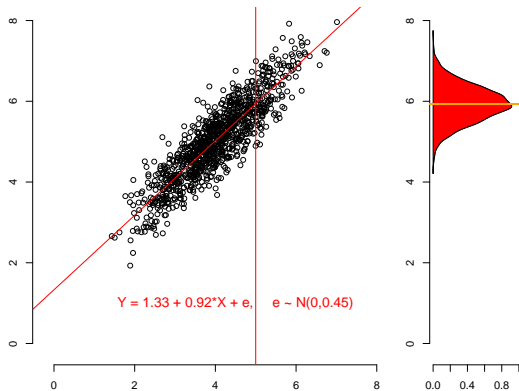
Y = 1.33 + 0.92*X + e,   e ~ N(0,0.45)

Example

What if instead of just observing genes and phenotypes, I was to *manipulate*/ *intervene on* / *change* the expression of that gene.

- E.g., deactivating or suppressing gene expression entirely.
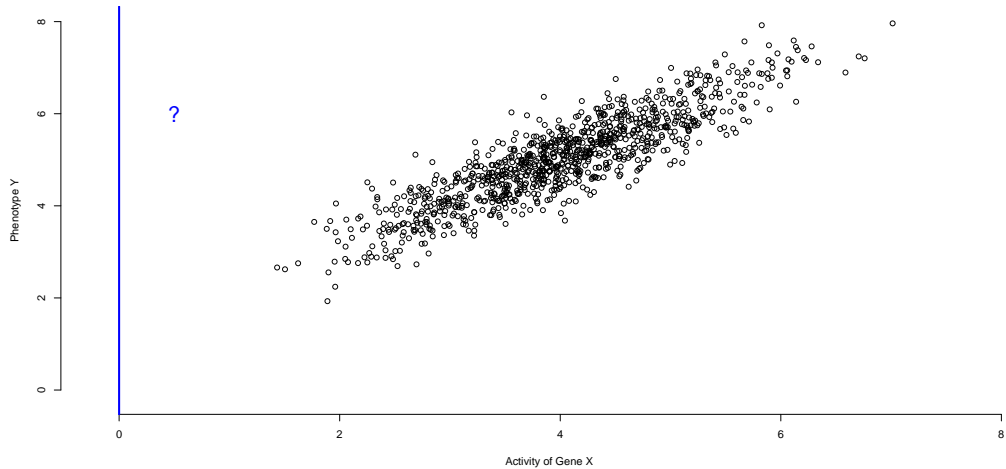
Example

Example

What if instead of just observing genes and phenotypes, I was to *manipulate/ intervene on / change* the expression of that gene.

- E.g., deactivating or suppressing gene expression entirely.

This is a research question which requires **causal reasoning** to resolve

- Predicting phenotype from gene expression in a *different setting*: The intervention setting instead of the observational setting
- Reasoning about the effects of an intervention *which we did observe*
- Reasoning about changing or interacting with the world

Example: Causal Reasoning
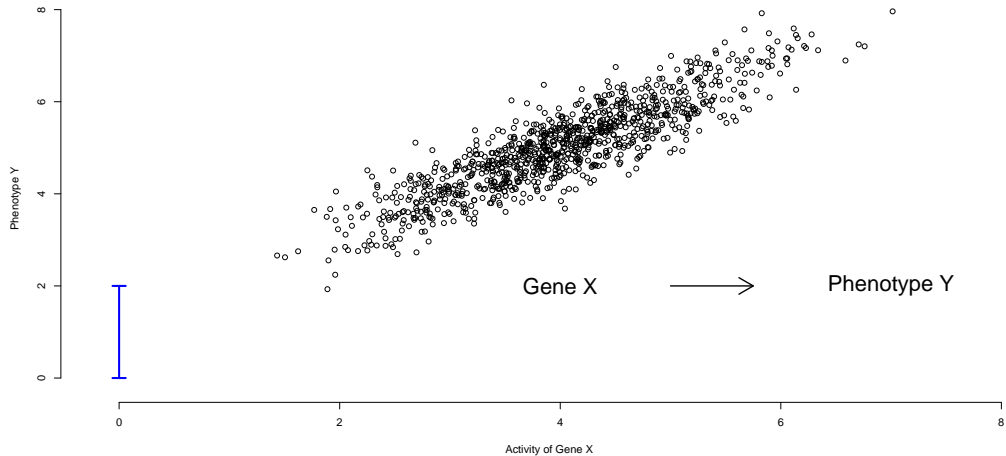
Example: Causal Reasoning
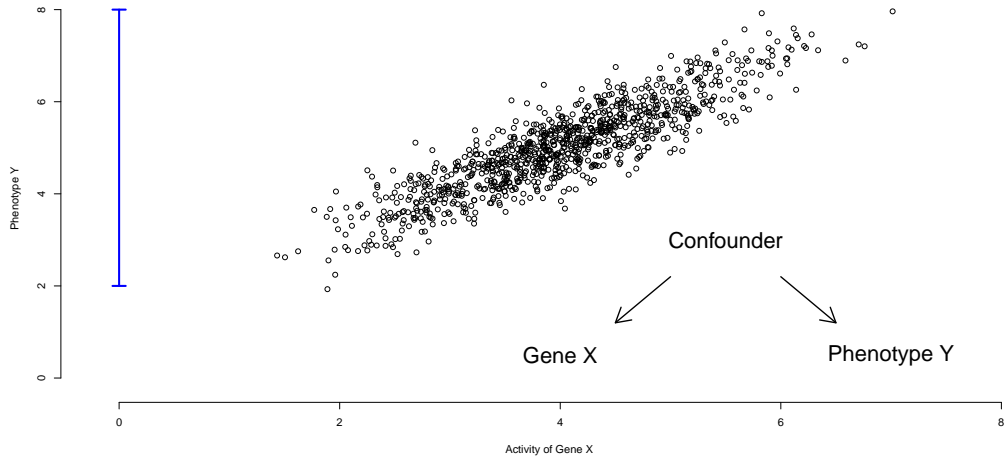
Statistical vs Causal Modeling

Statistical modeling and **data science** give us a rich language to describe uncertainty in the world we see around us

- The language of *co-occurences*, *expected values*, *(joint, marginal and conditional) probabilities* and *statistical dependencies*.
- It helps us *describe* patterns and make (certain types of) *predictions*.

Statistical vs Causal Modeling

Statistical modeling and **data science** give us a rich language to describe uncertainty in the world we see around us

- The language of *co-occurences*, *expected values*, *(joint, marginal and conditional) probabilities* and *statistical dependencies*.
- It helps us *describe* patterns and make (certain types of) *predictions*.

But *by themselves*, statistical models have very little to say about causal relations!

Statistical vs Causal Modeling

Causal Modeling involves using concepts and techniques from statistical modeling and data science

- But causal models and causal information exist on a level **above** statistical information

Causal modeling is a way of understanding when, how and which type of causal effects can be estimated from a particular dataset

- Involves reasoning about the subject matter (i.e. the real world) and how the dataset at hand was collected, amongst other factors
- Motivates/guides the application of particular types of statistical techniques

Causal models are not a special class of statistical techniques which magically guarantee that you can estimate causal effects from any dataset you happen to have

Randomized Control Trials

Randomized Control Trials (RCTs) are the gold standard for estimating causal effects.
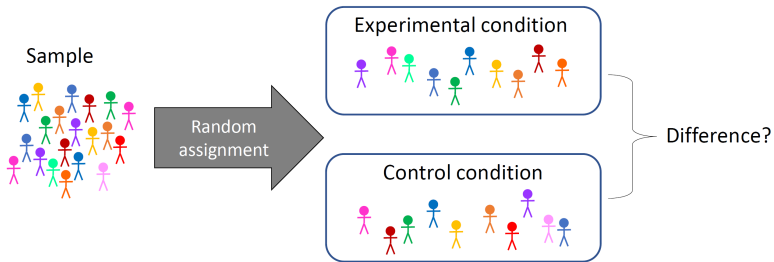
Randomized Control Trials

Randomized Control Trials (RCTs) are the gold standard for estimating causal effects.
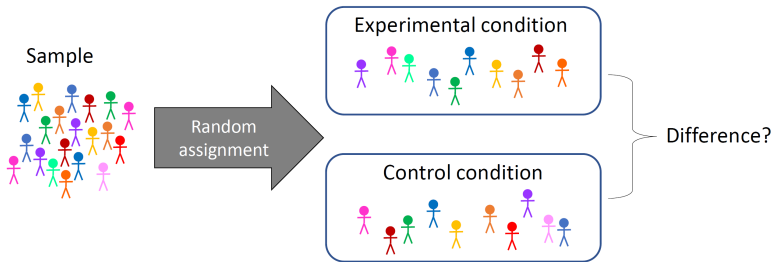
Great! But:

- What if the RCT doesn't work perfectly? What if I have non-compliance? Or if the intervention takes place over time?
- What if I can't perform an RCT due to ethical or practical constraints? Observational / non-experimental data?

How can we perform causal modeling in practice?

Temptation to split the world up into **two categories**:

- Randomized Experiments: We can estimate and talk about causal effects
- Non-randomized experiments, observational data: Not a randomized experiment, so don't even discuss causal relations

This conflates the **means** of research with the **ends** (Hernan, 2018)

This viewpoint muddies the waters

- In reality, the goal of much research is to learn about causal effects or causal relations
- Policing of causal language in observational studies doesn't change their goals, but leads to the use of euphemisms ("predicts", "relates", "is a risk factor for")
- Leads to confusion, poor methods, and a confused literature

How can we move forward?

Randomized experiments **are** special, but **why** are they special?

- What are the features, mechanisms, and principles by which randomized experiments typically lead to reliable causal inference?

By understanding these principles, can we understand if other designs might allow us to make the **same types of inferences**

- How can **mimic** those mechanisms, and **when can we make** the same types of inferences from other types of (e.g. observational) data?

We need a **language** to describe and understand causal inference

Two Frameworks / Languages for Causal Inference

Potential Outcomes (Today).

- Developed by statistician Don Rubin (m)
- Imbens (l) & Angrist (r): Nobel Prize for Economics 2021

Structural Causal Models / DAGs (Tomorrow).

- Developed by Judea Pearl, a computer scientist, amongst many others
- "Bayesian Networks"

Potential Outcomes I

Potential Outcomes

Headaches and Aspirin
- action: Aspirin ($X = 1$) or No Aspirin ($X = 0$)
- outcome: Headache gone ($Y = 1$) or Headache remains ($Y = 0$)

We want to know: Should I take an aspirin?
- I want to take aspirin if my headache level after taking aspirin is different than my headache levels if I don't take aspirin
- Two **potential versions of the outcome** for every person. Outcome if treated ($Y^{X=1}$) and outcome if not treated ($Y^{X=0}$)

A causal effect is defined as a **difference in potential outcomes**

Causal Effects

Individual Treatment Effect (or; Individual **Causal** Effect):

$$ITE_i = Y_i^{X_i=1} - Y_i^{X_i=0}$$

Causal Effects

Individual Treatment Effect (or; Individual **Causal** Effect):

$$ITE_i = Y_i^{X_i=1} - Y_i^{X_i=0}$$

The fundamental problem of causal inference (Holland, 1986): We can only observe one potential outcome per unit

Causal Effects

Individual Treatment Effect (or; Individual **Causal** Effect):
$$ITE_i = Y_i^{X_i=1} - Y_i^{X_i=0}$$

The fundamental problem of causal inference (Holland, 1986): We can only observe one potential outcome per unit

If you decide to take the aspirin $(x = 1)$, in this situation I will observe your headache outcome under aspirin-taking: $Y_i^{X_i=1}$

- This is sometimes referred to as the **factual** outcome

But that means I cannot observe your headache outcome under aspirin-avoidance: $Y_i^{X_i=0}$

- This is then referred to as your **counterfactual** outcome

Causal Effects and Missing Data

From the potential outcomes perspective, causal inference is a **missing data problem**

How can we approach solving this problem?

On a "meta" level, there are two basic strategies we can (and must) engage in:

1. Change the **estimand**: Ask a different causal question, one that might be easier to answer

2. Make **assumptions**: Use information and knowledge about the real world we do have to make informed guesses about information we don't have

On Causal (Identifiability) Assumptions

The topic of **assumptions** is central to all approaches to causal modeling you will learn about this week

Causal Identifiability Assumption:

What would have to be true in the context in which this dataset was collected for us, in theory, to be able to get at the causal effect we are interested in (assuming infinite data, and without reference to any estimation approach)

We will learn about **three generic** causal identifiability assumptions in the potential outcomes framework

They are not the same as statistical assumptions (e.g., heteroskedasticity, linearity, etc.), and typically cannot be verified empirically.

Which Estimand?

ITE estimation is generally quite tricky

- the assumptions we need to identify it are very strong or unlikely to hold in many practical scenarios

Instead we typically focus on trying to infer the **average treatment effect**

> **Average Treatment Effect** (or Average **Causal** Effect):
> $$ATE = E[Y_i^{X=1} - Y_i^{X=0}] = E[Y^1] - E[Y^0]$$

More feasible to estimate (i.e. typically we can make fewer or more **realistic** assumptions) when we have many observations from different individuals, and often sufficient for many practical decision making situations

Example: Aspirin and Headaches

| | Potential outcomes | | ICE |
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$
Charles	1	1	0
George	0	0	0
Susan	1	0	1
Tracy	1	1	0
Ken	0	1	-1
Pete	1	0	1
Helen	1	0	1
Kate	0	0	0

Example: Aspirin and Headaches

| | Potential outcomes | | ICE |
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$
Charles	1	1	0
George	0	0	0
Susan	1	0	1
Tracy	1	1	0
Ken	0	1	-1
Pete	1	0	1
Helen	1	0	1
Kate	0	0	0

$$ATE = E[Y^1] \quad - \quad E[Y^0]$$
$$ATE = 5/8 \quad - \quad 3/8 \quad = 0.25$$

But we only observe one outcome per person

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	1	1	0	0	1
George	0	0	0	1	0
Susan	1	0	1	1	1
Tracy	1	1	0	0	1
Ken	0	1	-1	0	1
Pete	1	0	1	1	1
Helen	1	0	1	0	0
Kate	0	0	0	1	0

But we only observe one outcome per person

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	1	1	0	0	1
George	0	0	0	1	0
Susan	1	0	1	1	1
Tracy	1	1	0	0	1
Ken	0	1	-1	0	1
Pete	1	0	1	1	1
Helen	1	0	1	0	0
Kate	0	0	0	1	0

Expected value of recovery **aspirin takers** ($X = 1$): $(0+1+1+0)/4 = 0.5$
Expected value of recovery **aspirin avoiders** ($X = 0$): $(1+1+1+0)/4 = 0.75$

But we only observe one outcome per person

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	1	1	0	0	1
George	0	0	0	1	0
Susan	1	0	1	1	1
Tracy	1	1	0	0	1
Ken	0	1	-1	0	1
Pete	1	0	1	1	1
Helen	1	0	1	0	0
Kate	0	0	0	1	0

Expected value of recovery **aspirin takers** ($X = 1$): $(0+1+1+0)/4 = 0.5$
Expected value of recovery **aspirin avoiders** ($X = 0$): $(1+1+1+0)/4 = 0.75$

$E(Y|X = 1) - E(Y|X = 0) = -0.25$

Naive conclusion: Aspirin decreases chances of headache relief.

What is the problem with observing?

Observing $\neq$ intervening

$E(Y|X = 1) - E(Y|X = 0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X = 1) \neq E(Y|X = 0)$ (in words: the average value of headache levels for those who did and did not take aspirin are unequal), does not, in general, **imply a causal effect** of X on Y.

What is the problem with observing?

Observing $\neq$ intervening

$E(Y|X=1) - E(Y|X=0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X=1) \neq E(Y|X=0)$ (in words: the average value of headache levels for those who did and did not take aspirin are unequal), does not, in general, **imply a causal effect** of X on Y.

In RCTs, we often use $E(Y|X=1) - E(Y|X=0)$ as an **estimate** of the *ATE*.

What is the problem with observing?

Observing $\neq$ intervening

$E(Y|X=1) - E(Y|X=0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X=1) \neq E(Y|X=0)$ (in words: the average value of headache levels for those who did and did not take aspirin are unequal), does not, in general, **imply a causal effect** of X on Y.

In RCTs, we often use $E(Y|X=1) - E(Y|X=0)$ as an **estimate** of the *ATE*. But why? What makes an RCT so special?

Assumption 1: Exchangeability

At best, **half** of the potential outcomes are **observed**; hence, causal inference is at its core a **missing data problem**.

The critical question is: What is the **missing data mechanism**?
Or: What is the **assignment mechanism**?

If there is a relation between the **assignment mechanism** and the **potential outcomes**, this may bias the estimation of the causal effect.

Exchangeability:

The actual treatment received (X) and the potential outcome given treatment Y^X are independent: $Y^x \perp\!\!\!\perp X$ for all x

This is also known as **unconfoundedness**: The missing potential outcome is missing **completely at random**. Individuals across treatment groups are **exchangeable**

Treatment NOT independent of potential outcomes

In a **non-randomized study**, treatment may depend on person features that also relate to the potential outcomes.

	Unobserved			Observed		Confounder
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i	Z_i
Charles	1	1	0	0	1	3
George	0	0	0	1	0	9
Susan	1	0	1	1	1	8
Tracy	1	1	0	0	1	5
Ken	0	1	-1	0	1	4
Pete	1	0	1	1	1	2
Helen	1	0	1	0	0	5
Kate	0	0	0	1	0	4

Average headache levels are higher among those who took the aspirin. But, people who took aspirin also scored higher on the covariate **dehydration levels** Z_i.

Conditional Exchangeability

Luckily, we don't need full exchangeability for causal inference. We only need **conditional exchangeability**; conditional on a set of observed covariates, the potential outcomes are independent of treatment assignment.

Conditional exchangeability:

The actual treatment received (X) and the potential outcome given treatment Y^X are independent within certain levels of Z: $Y^x \perp\!\!\!\perp X|Z$

This implies that data are *missing at random* (rather than *missing completely at random*).

Estimation of the ATE can proceed **as long as we can properly account for (i.e. condition on) the confounder** Z. That is, if all **confounders** are **observed** and we can **control** for them. But to be able to do this, we need...

Assumption 2(a): Consistency

> **Consistency**:
>
> For each unit, we observe one of the potential outcomes of interest: Y^1 for individuals with $X = 1$ and Y^0 for those with $X = 0$.
>
> Treatment is **well-defined**: there are **not different versions of each treatment level** that lead to different potential outcomes.

Consistency ensures that we are making inferences about an actual target intervention.

- Subtle assumption that is **very often violated** in observational settings

Different treatments

If there are multiple ways to raise X from 0 to 1, this means:

- there are **multiple treatments**
- these may have **different effects**
- and hence the causal question is **ill-defined**

Examples:

- What is the effect of obesity on health?
- Does physical punishment compromise children's well-being?
- Does alcohol undermine cognitive performance in young adolescents?

To formulate better questions, we should define the **target trial**: The randomized controlled trial we would have done, if it had been possible.

Assumption 2(b): No Interference

No Interference:

The potential outcomes for any unit do not vary with the treatments assigned to other units

E.g., we are interested in the effects of ritalin on concentration levels among students in a classroom. Alice may be better able to concentrate because Bob takes ritalin and disrupts the classroom less

- Alice's value of Y^0 depends on Bob's value of X

Putting it together: SUTVA

Stable unit treatment value assumption (SUTVA):

The potential outcomes for any unit do not vary with the treatments assigned to other units (i.e., **no interference**), and,
for each unit, there are **not different versions of each treatment level** that lead to different potential outcomes.

Assumption 3: Positivity

There must be **exposed and unexposed participants** at every combination of values of Z in the population under study.

In an **RCT**, positivity is **present by design** (in the expectation)

In a **non-experimental study**, **violations** can be **detected** by:

- making tables of each categorical covariate and treatment (should be no empty cells)
- categorize a continuous covariate and make table (but this depends on number and width of categories)
- considering all combinations of covariates (becomes impossible)

Putting it Together

Three Conditions/Assumptions necessary for causal **identification**:

1. (Conditional) Exchangeability
2. SUTVA (consistency and no interference)
3. Positivity

Causal Estimation:
Given our data and causal identification assumptions, how should we estimate the causal effect

Practical 1

Check-in exercise

In **groups**, discuss one of the following research questions and data settings. Which identifiability assumption(s) might be violated, and why?

1. What is the effectiveness of a single shot of the 2024 flu vaccine compared to no vaccine, on preventing respiratory illnesses in the population at large? We have data from GPs who administered the vaccine primarily in elderly care homes.

2. Does drinking coffee prevent bowel cancer? We have information from a randomly selected panel/cohort, where we track their coffee consumption and any cancer diagnoses over 10 years.

3. Does taking after-school classes improve educational outcomes? We have data from different schools where voluntary after-school classes were offered.

Estimation

The Two "Tasks" of Causal Inference

Identification

Assuming I have **population-level statistical information** (given these variables but with an infinite sample size), can I infer the causal effect of interest?

What causal assumptions/conditions need to be met?

Estimation

Given that my causal effect is identified, how should I go about estimating this effect from sample data?

Statistical assumptions - functional form, distributions, etc.

Causal Inference and Estimation

Generally in causal inference settings the aim is:

1. Mimic "randomization" using statistical tools: **adjustment** approaches
2. Do so by making as few additional statistical assumptions as possible (towards non-parametric methods)

Statistics in a nutshell

Statistical Estimand

Estimator

Estimate

1. **Prepare Chocolate Cake Batter**
   Preheat oven to 350 degrees, and prepare Yo's Ultimate Chocolate Cake batter. Prepare your pans with parchment. Pour 2 ½ lbs into each 7" round pan, 1 ½ lbs into your 6" round pan, and divide the remaining batter evenly between your 5" round pans.

2. **Bake Cakes**
   Bake your 7" round cakes for 50 minutes, your 6" round cake for 40 minutes, and your 5" round cakes for 30 minutes, or until a toothpick comes out clean. Set aside to cool completely in their pans on a wire rack.

3. **Prepare Fillings & Simple Syrup**
   Prepare your dark chocolate ganache, Italian meringue buttercream, and simple syrup. Set aside until you're ready to decorate.

4. **Level Cakes**
   Remove your cooled cakes from their pans and level them with a ruler and serrated knife.

Causal Inference in a nutshell

Causal Estimand	**Causal Model**	**Statistical Estimand**	**Estimator**	**Estimate**

Causal Inference in a nutshell

Causal Estimand	**Causal Model**	**Statistical Estimand**	**Estimator**	**Estimate**

Conditional Exchangeability by Conditioning on Confounders

We want to **condition** on or **adjust** for variables that affect treatment assignment and (potentially) the outcome of interest: confounders

When conditioning, we want to ensure that, within fixed levels of the confounders, the control and treatment groups are **conditionally exchangeable**

There are many many different ways to condition on a variable(s)

- you are probably already familiar with regression models!

How to choose confounders? Must be based on background knowledge: we will return to this tomorrow and Wednesday

Treatment NOT independent of potential outcomes

In a **non-randomized study**, treatment may depend on person features that also relate to the potential outcomes.

	Unobserved			Observed		Confounder
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i	Z_i
Charles	1	1	0	0	1	3
George	0	0	0	1	0	9
Susan	1	0	1	1	1	8
Tracy	1	1	0	0	1	5
Ken	0	1	-1	0	1	4
Pete	1	0	1	1	1	2
Helen	1	0	1	0	0	5
Kate	0	0	0	1	0	4

Average headache levels are higher among those who took the aspirin. But, people who took aspirin also scored higher on the covariate **dehydration levels** Z_i.

Adjustment by Stratification

1. Define strata or levels of the covariate(s) of interest. E.g.,
   - $Z = 0$, $Z = 1$
2. Estimate the group difference within those strata.
   - $E[Y|X = 1, Z = 1] - E[Y|X = 1, Z = 1]$
   - $E[Y|X = 1, Z = 0] - E[Y|X = 1, Z = 0]$
3. Take the weighted average, weighted by the number of people in each strata
   - $(E[Y|X = 1, Z = 1] - E[Y|X = 1, Z = 1])p(Z = 1)$
4. Take the average to obtain the ATE/ACE
   - $\hat{ATE} = \sum(E[Y|X = 1, Z = z] - E[Y|X = 0, Z = z])p(Z = z)$

Adjustment by matching

Conceptually similar to adjustment by stratification

For every person in your dataset, find someone with the same set of covariate values

This enforces that covariates are **balanced** cross groups

The difference between matched treated and untreated groups is an estimate of the ATE

More in the practical!

Propensity Scores

Propensity Scores are a tool used primarily for causal estimation

Propensity scores (assuming no unobserved confounding):

The probability of exposure/treatment given confounders Z
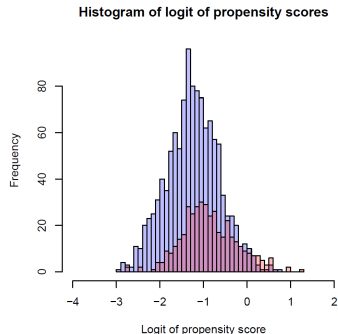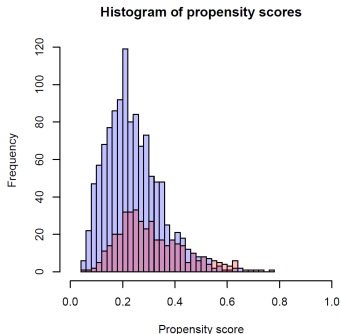$$\pi_i = P[X_i = 1 | Z_i] = \frac{exp(Z_i'\phi)}{1 + exp(Z_i'\phi)}$$

We estimate π_i using **logistic regression**

Propensity scores:

- Summarize information about the relationship between **pre-treatment** confounders (Z) and treatment (X)
- Are used to ensure *conditional exchangeability*

Get $Y^X \perp\!\!\!\perp X | \pi$ to replace $Y^X \perp\!\!\!\perp X | Z$

Overlap of propensity scores

Histogram of propensity scores

Histogram of logit of propensity scores

The **distributions of logit**(π_i) for the treated and the untreated are typically different, but should fully (and properly) **overlap**:

- non-overlapping areas imply **violation of positivity assumption**
- non-overlapping areas **require extrapolation**
- areas with very few people in one groups imply there are **few matches**

Propensity Scores for Matching

Matching implies you **create pairs** that consist of a treated and a non-treated person, who have **identical propensity scores**.

Background: In an **RCT** we have: $P(Z|X = 1) = P(Z|X = 0)$

Balancing property:
$$P(Z|\pi = c, X = 1) = P(Z|\pi = c, X = 0)$$

If the propensity model is **correct**, then comparing treated and untreated **individuals with the same π** is a way of **mimicking an RCT**.

Practical 2

Potential Outcomes: An Overview

Causal Inference is a missing data problem

- When can I infer $E[Y^1] - E[Y^0]$ if i don't fully observe either?

Steps (broadly):

- Assess SUTVA, Exchangeability and Positivity
- If you can meet those conditions, use covariate-based techniques like matching/stratitification (with or without propensity scores) to adjust for/create balanced groups of treated and not treated, mimicing what would occur by design in an RCT
- Estimate ATE by adjusting for group differences on confounders

Recommended Reading

Hernán, M. A. (2018). The C-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, 108(5), 616-619.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27-42.

Schafer, J. L., Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4), 279.

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman Hall/CRC. Free copy: `https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf`

Pearl, J., Glymour, M. & Jewell, N.P. (2016) Causal Inference in Statistics: A Primer