Introduction to Causal Inference and Causal Data Science

Day 1: Causal Inference and Potential Outcomes

Oisín Ryan

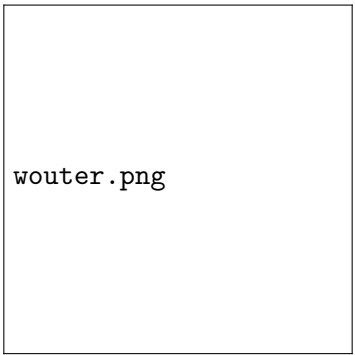
Department of Data Science and Biostatistics
Julius Center
UMC Utrecht

July 1, 2024

Slide on participants; backgrounds / logo's from different universities / disciplines

Your teachers

Oisín Ryan



Wouter van Amsterdam

Bas Penning de Vries

Why Causal Modeling?

Causal questions dominate scientific research

- Does smoking cause cancer?
- 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?
- Which treatment type will be most effective in reducing symptoms for this type of individual?

Why Causal Modeling?

Causal questions dominate scientific research

- Does smoking cause cancer?
- 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?
- 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?

Why 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*.

Why 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!

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

Example

Imagine we are a team of health scientists.

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

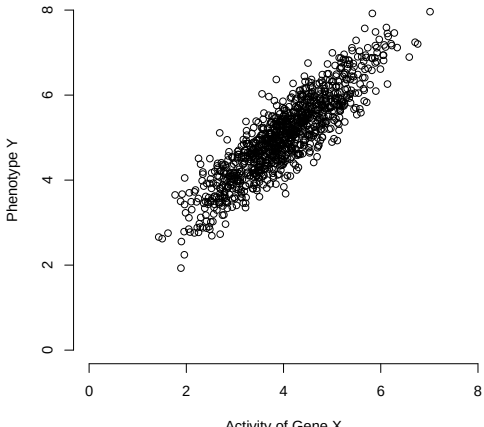
- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).

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).

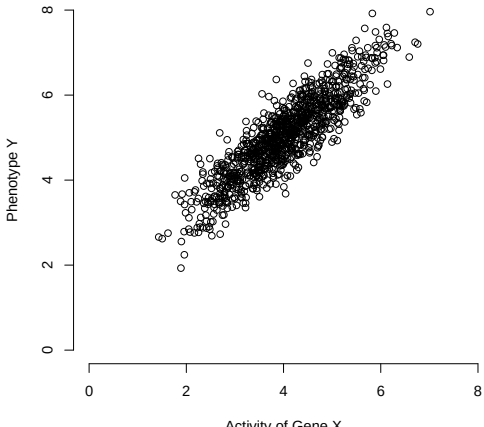


Example

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).

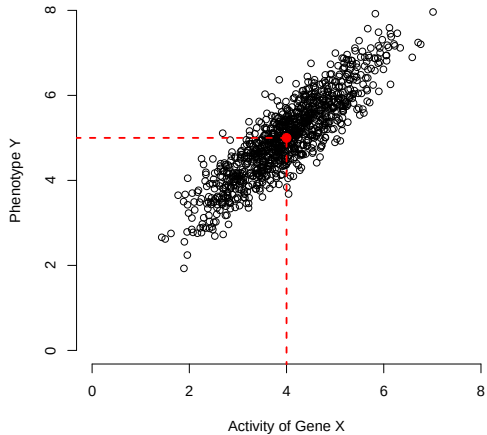


What kind of information can we extract from this data? What tasks can we perform, and what research questions can we answer using statistical techniques?

Example

Description:

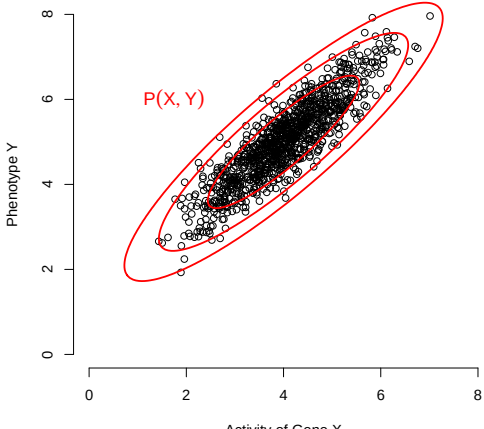
- What is the average level of gene expression in the sample ($\bar{X}$)?
- What does that tell us about the average level in our population ($E[X]$)?
- How *certain* are we about our *estimate* of the population mean?



Example

Models for (co-)occurence

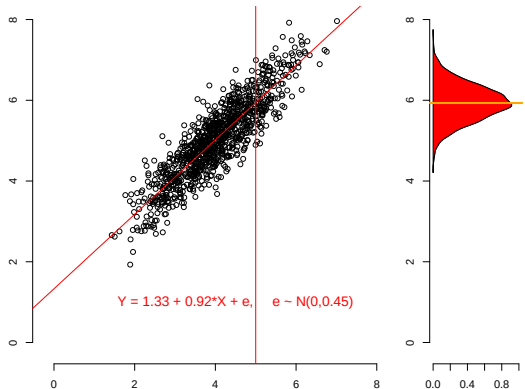
- What is the likelihood of observing a low level of gene expression (*Marginal Distribution* $P(X = x)$)?
- What is the probability that someone in the population has both a high insulin level and a high gene expression (*Joint distribution* $P(X, Y)$)?
- We can fit models, such as the normal distribution $P(X, Y) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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



Example

What kinds of questions can we **not answer**?

Example

What kinds of questions can we **not answer**?

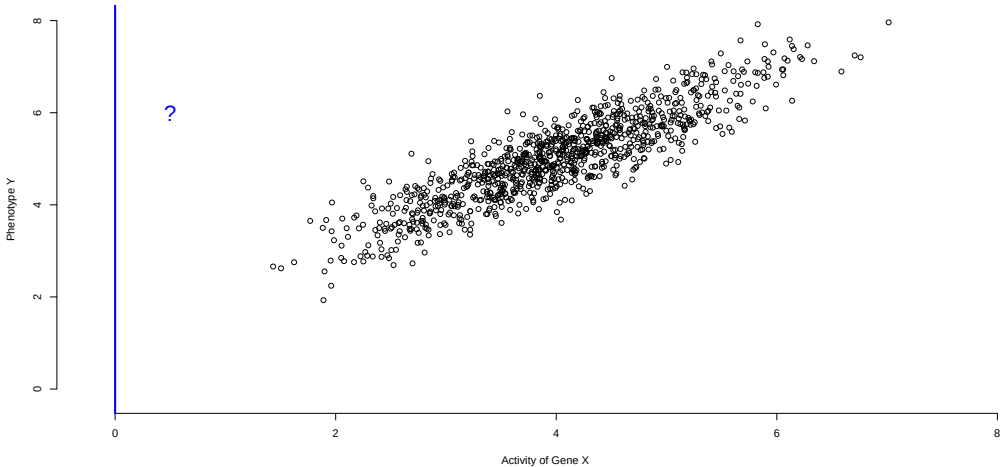
What if instead of just observing genes and phenotypes, I was to *manipulate*/ intervene on / *change* the expression of that gene.

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

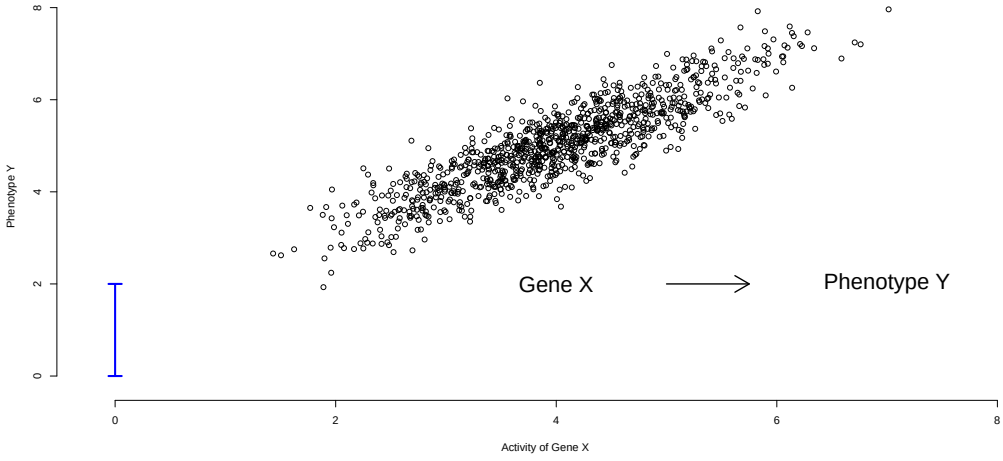
What level of phenotype expression would I expect to see if I did that?

- Predicting phenotype from gene expression in a different setting: The intervention setting instead of the observation setting

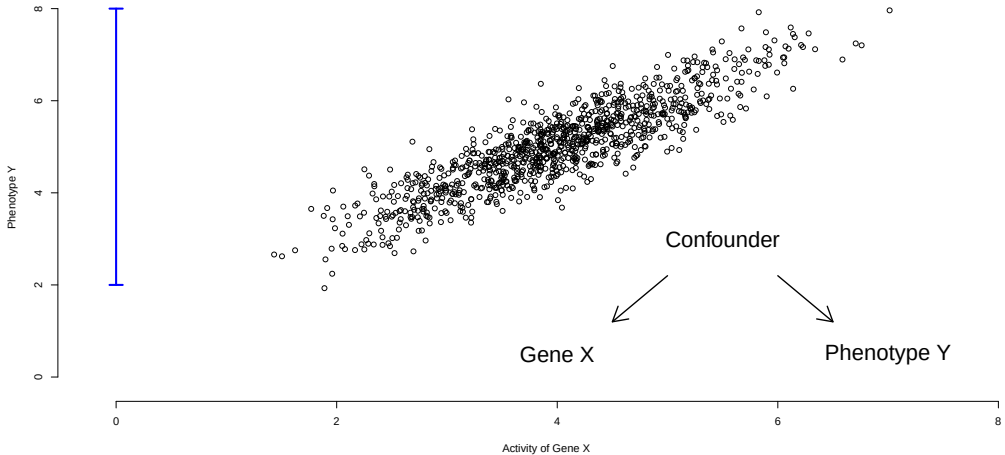
Example: Causal Reasoning



Example: Causal Reasoning

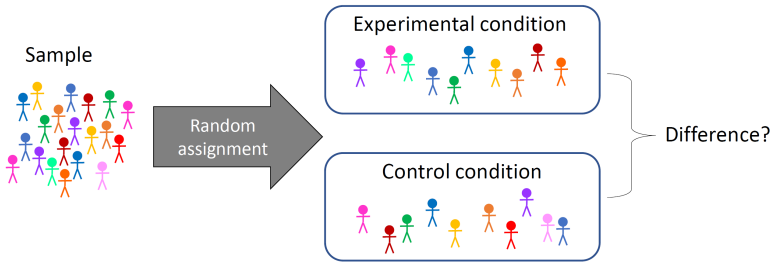


Example: Causal Reasoning



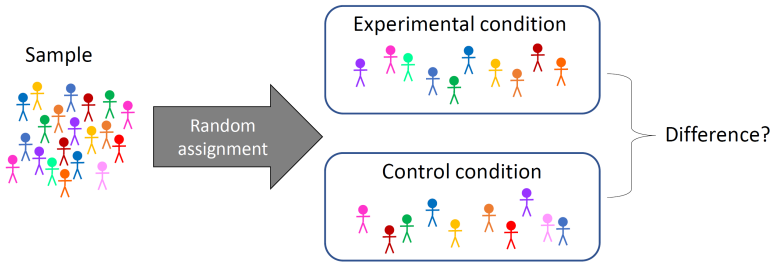
Randomized Control Trials

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?
- 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? (rough)

Temptation to split the world up into two dimensions:

- 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* with the *ends* (Hernan citation)
Leads to confusion; in reality, many causal reserach questions cannot or cannot efficiently/effectively/practicaly be answered using randomized experiments. - euphemisms, confusion

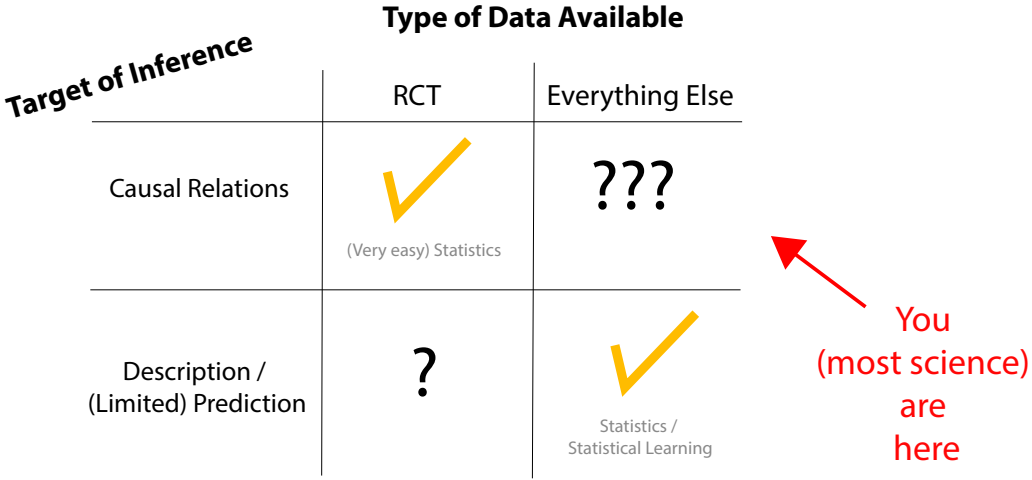
Science: A 2x2 table

Target of Inference	Type of Data Available: RCT	Type of Data Available: Everything Else
Causal Relations		
Description / (Limited) Prediction		

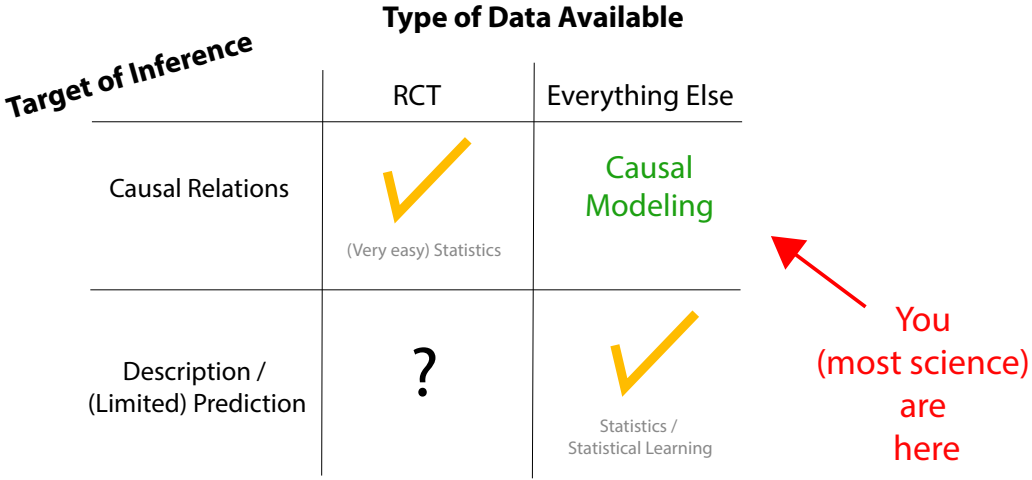
Science: A 2x2 table

Target of Inference	Type of Data Available: RCT	Type of Data Available: Everything Else
Causal Relations	✓ (Very easy) Statistics	???
Description / (Limited) Prediction	?	✓ Statistics / Statistical Learning

Science: A 2x2 table



Science: A 2x2 table



How can we perform causal modeling in practice? (rough)

Randomized experiments **are** special, but **why** are they special?
- what are the features / mechanisms / principles by which this study design allows for causal inference?
- by understanding these, can we understand if other designs might allow us to make the same types of inferences, e.g., by mimicing / accounting for those mechanisms etc.
We need a language to describe and understand causal inference

Two Frameworks / Languages for Causal Inference

Potential Outcomes (Part 1).

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

Structural Causal Models (Part 2).

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

Outline

Outline

Structure of the course etc

Lecture, then lab
Lunch
(schedule for today)

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 causal effect:

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

Causal Effects

Individual causal effect:

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

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

Causal Effects

Individual causal effect:

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

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

- Highlights the basic structure of the causal inference problem from this perspective
- We have a missing data problem
- How can we approach this?
- Two "knobs" you can turn; 1) change the question you're asking, and 2) use information you do have to make guesses about the information you don't have
- In (2) the most crucial step is to understand in what situations we can use **information we do have** to make correct/useful/helpful guesses - These descriptions of "situations" are known as **assumptions**

Which causal effect?

Let's turn both knobs here; turns out ICE estimation is quite tricky in most situations - Instead it is often more feasible (and often of primary interest anyway) to estimate Average causal effects

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

Average causal effect:

$$ACE = E[Y_i^{X=1} - Y_i^{X=0}] = E[Y^1] - E[Y^0]$$

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

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

$$ACE = E[Y^1] - E[Y^0]$$
$$ACE = 5/8 - 3/8 = 0.25$$

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 *ACE*.

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 *ACE*. But why? What makes an RCT so special?

NOTES

Rearrange below to make exchangeability first,
Split up SUTVA into no interference and consistency

Assumption 1: 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.

SUTVA is important when moving from a single unit (with an ICE), to multiple units (when we consider the ACE).

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: 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 ACE can proceed **as long as we can properly account for (i.e. condition on) the confounder** Z. But to be able to do this, we need...

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 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. SUTVA
2. (Conditional) Exchangeability
3. Positivity

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

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.

Philosophy on causal inference and estimation:
- Make as few additional assumptions as possible - Mimic "randomization" using statistical tools; towards "non-parametric" **adjustment** approaches
Recall: we want to "adjust" for variables that determine which treatment your receive and which value of the outcome you are likely to have

Statistics in a nutshell

Statistical Estimand

Estimator

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.

Estimate

Causal Inference in a nutshell

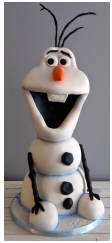
Causal Estimand	Causal Model	Statistical Estimand	Estimator	Estimate

Causal Inference in a nutshell

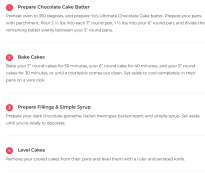
Causal Estimand

Causal Model

Statistical Estimand



Estimator



Estimate

Conditional Exchangeability by Conditioning

Recall: Conditional exchangeability
Within levels of the confounder, groups are comparable
Conditionally randomized trial

Adjustment by Stratification

- define strata ($Z = 0$, $Z = 1$)
- Estimate average effect within those strata
- ACE = weighted sum of these (more dehydrated than not, etc.)

Adjustment by matching

- like stratification taken to an extreme
- for every person in your dataset, find someone with the same set of covariate values
- take the difference in outcome in each matched pair
- the average of these = ACE

Propensity Scores

Propensity Scores are a tool used in the PO framework for causal estimation

Propensity scores (assuming no unobserved confounding):

The probability of exposure/treatment given confounders Z

$$\pi_i = P[X_i = 1|Z_i] = \frac{exp(Z_i'\phi)}{1+exp(Z_i'\phi)}$$

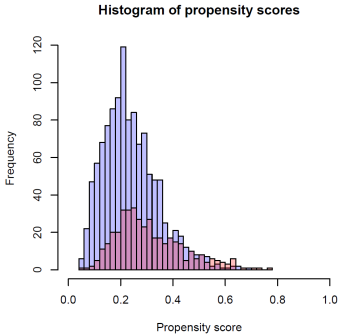
We estimate π_i using **logistic regression**

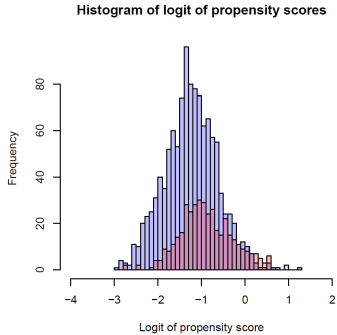
Propensity scores:

- 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





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: $P\big(Z|X=1\big) = P\big(Z|X=0\big)$

Balancing property:

$$P\big(Z|\pi=c, X=1\big) = P\big(Z|\pi=c, X=0\big)$$

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

Propensity Scores for Weighting

Check if this will be covered in Day 3 (probably will be) The **probability of received treatment** is:

- π_i for those who were **treated** $(X_i = 1)$
- $1 - \pi_i$ for those who were **NOT treated** $(X_i = 0)$

Among **treated individuals**, those with large π_i are **overrepresented** in comparison to those with small π_i.

Among **untreated individuals**, those with large $1 - \pi_i$ are **overrepreseneted** in comparison to those with small $1 - \pi_i$.

To account for this inbalance, we **create a pseudo-population** where each case is **weighted** by the **inverse probability of received treatment**:

- $\frac{1}{\hat{\pi}_i}$ for $X_i = 1$
- $\frac{1}{1-\hat{\pi}_i}$ for $X_i = 0$

Causal Estimands

ATE and ATT
Matching/Weighting and ATE/ATT

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 propensity scores to create balanced groups of treated and not treated, mimicing an RCT
- Estimate ACE by adjusting for group differences on confounders (e.g., weighting, matching)

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