Causal Inference: Methods for Program Evaluation and Policy Research

Prof. Jennifer Hill
Fall 2018

Acknowledgements

I would like to acknowledge the contributions to these slides and this course provided by

Andrea Cornejo

Zarni Htet

Rui Lu

in the form of intellect, creativity, humor, patience, and positive energy!

Everything you like about these slides is likely due to them. All errors are my own.

Learning objectives for today

1. What is causal inference?
    - Why is causal inference important?
    - How is it defined?
    - What is the difference between association and causation?
2. What are counterfactual outcomes?
    - Conceptually?
    - How do we encode notationally? Potential outcomes
    - What is the fundamental problem of causal inference?
3. How do we define causal effects?
4. What is the fundamental problem of causal inference?

What is causality?
Philosophical viewpoints

A causal relationship between an event and outcome, has been debated and defined in many fields.

(And continues to be debated even today….)

Philosophical views include...

"Something that makes a difference, and the difference it makes must be a different from what would have happened without it."

- David Lewis, American philosopher of the 20th century

"Causality is a genetic connection of phenomena through which one thing (the cause), under certain conditions, gives rise to causes something else (the effect)."

- Alexander Spirkin, Russian philosopher of the 20th century

What is causality?
Philosophical viewpoints

Statistician's views include…

"Correlation is not causation."

-R.A Fisher British Statistician of the 20th century

"True treatment effects in an experiment in almost exactly the same way that we have defined causal effects."

-David Cox , British Statistician of the 20th century

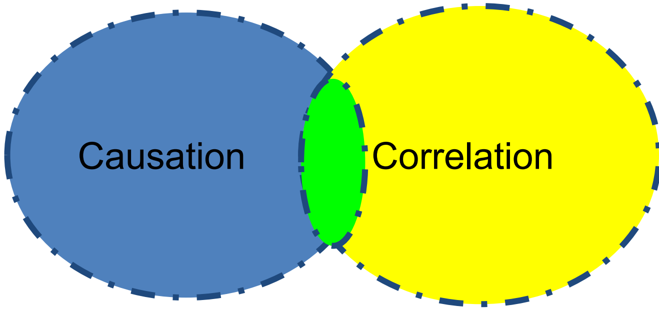
Causation is "the effects caused by treatments as comparison of potential outcomes under alternative treatments."

-Jerzy Neyman, Polish Statistician of 20th century

Correlation and Causation

Correlation is a statistical measure (expressed as a number) that describes the size and direction of an empirical relationship between two or more variables.

- Correlation measures the strength of the linear association between 2 or more quantitative variables.
- Positive correlation could be consistent with negative causal effects, positive causal effect or lack of causal effect.
- For example, taking cough medication is positively correlated with coughing : do we think that it causes coughing?
- Other examples…



Social Isolation example. Do the following:

"A wave of new research suggests social separation is bad for us. Individuals with less social connection have disrupted sleep patterns, altered immune systems, more inflammation and higher levels of stress hormones. One recent study found that isolation increases the risk of heart disease by 29 percent and stroke by 32 percent."

1) Draw a scatterplot of hypothetical data that match this description about the relationship between
   a) degree of social connection (imagine this on a continuous scale from 0 to 1)
   b) levels of stress hormones
2)

Social Isolation example. Do the following:

"A wave of new research suggests social separation is bad for us. Individuals with less social connection have disrupted sleep patterns, altered immune systems, more inflammation and higher levels of stress hormones. One recent study found that isolation increases the risk of heart disease by 29 percent and stroke by 32 percent."

1) Draw a scatterplot of hypothetical data that match this description about the relationship between
   a) degree of social connection (imagine this on a continuous scale from 0 to 1)
   b) levels of stress hormones
2) Compare your plot with your neighbor's and discuss any differences in the features (level of association, sign of association, shape of association, etc.)

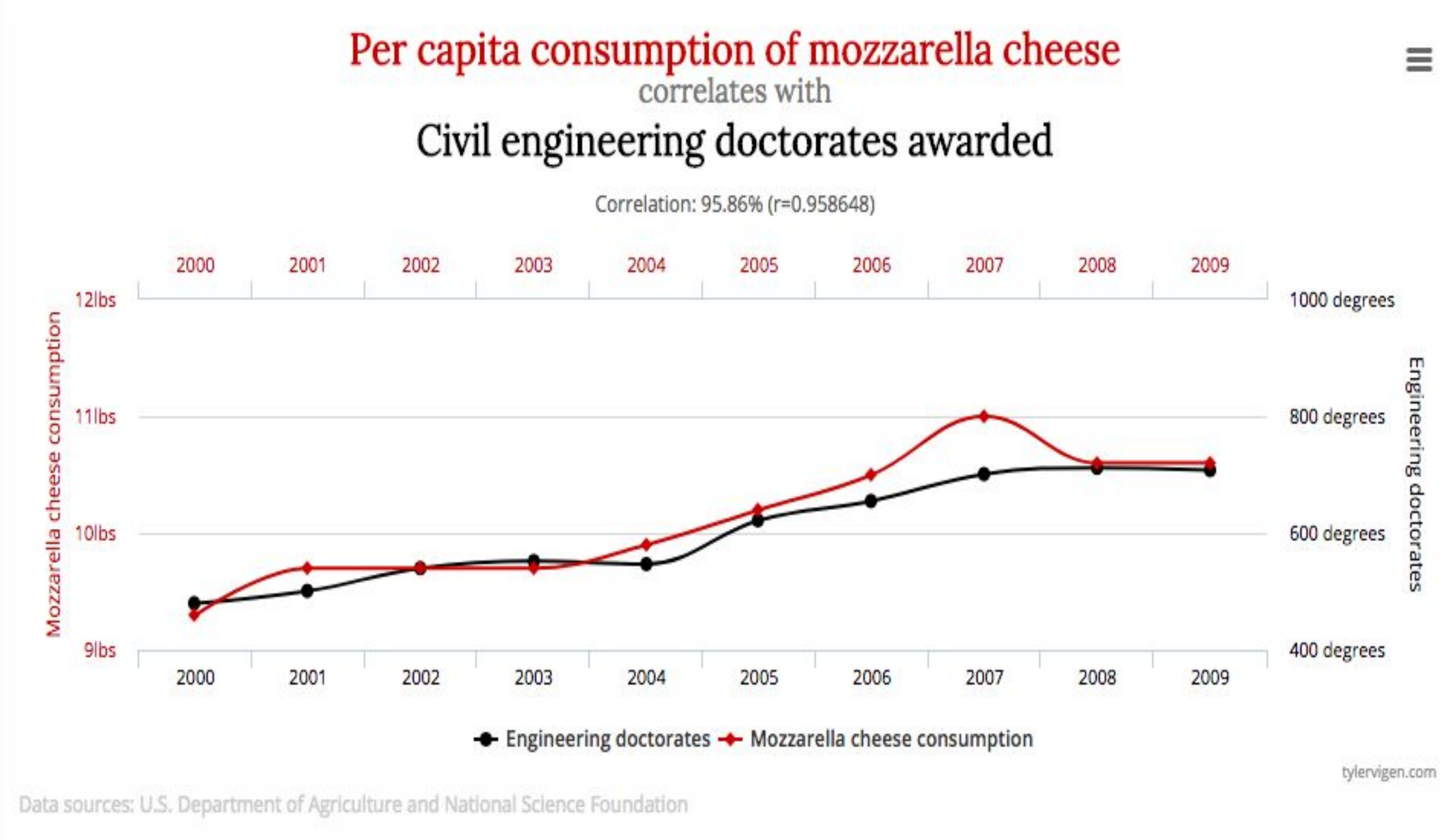
Social Isolation example. Do the following:

"A wave of new research suggests social separation is bad for us. Individuals with less social connection have disrupted sleep patterns, altered immune systems, more inflammation and higher levels of stress hormones. One recent study found that isolation increases the risk of heart disease by 29 percent and stroke by 32 percent."

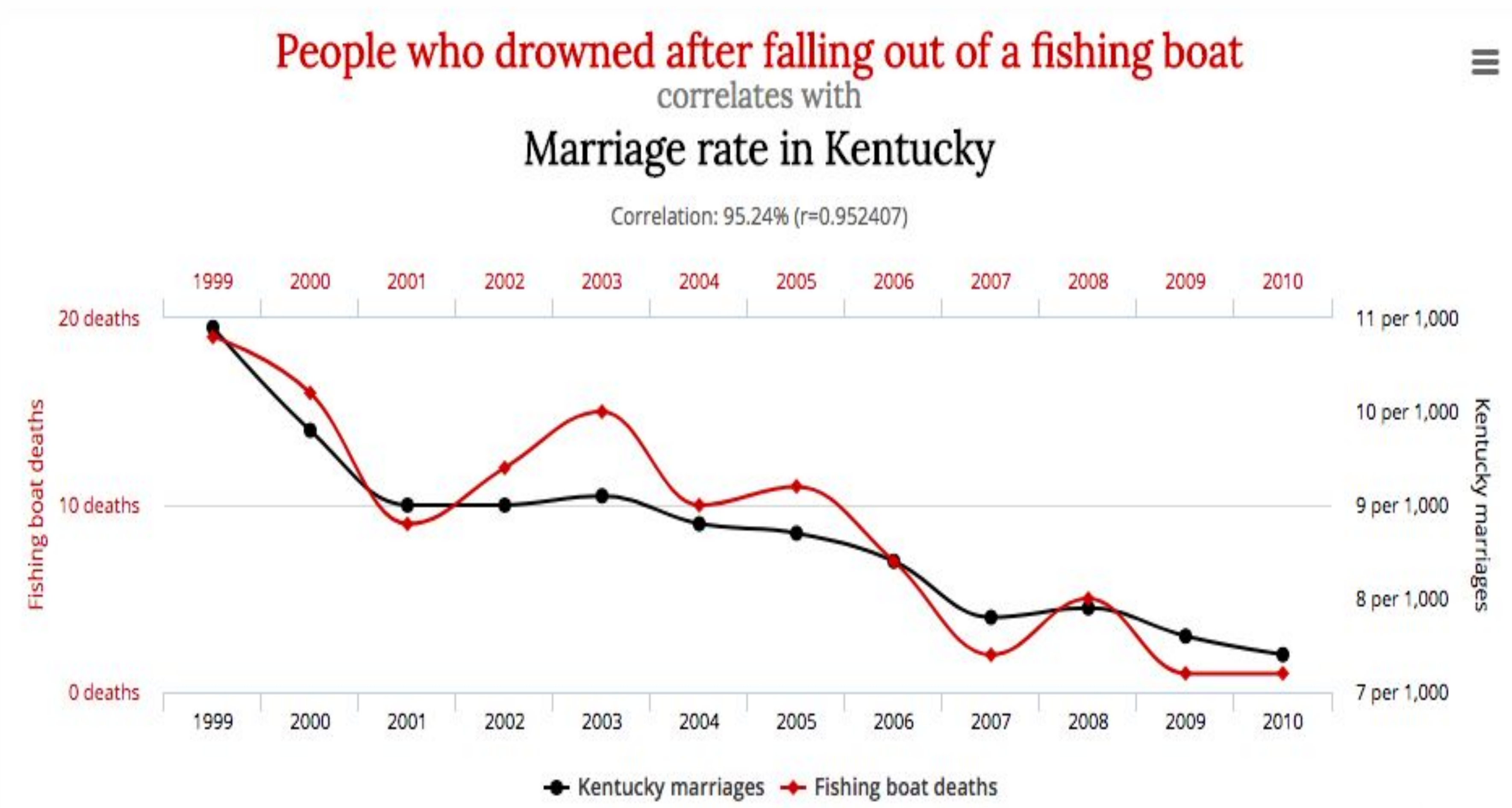
1) Draw a scatterplot of hypothetical data that match this description about the relationship between
   a) degree of social connection (imagine this on a continuous scale from 0 to 1)
   b) levels of stress hormones
2) Compare your plot with your neighbors and discuss any differences in the features (level of association, sign of association, shape of association, etc.)
3) Do you think this relationship is causal?

Correlation and Causation

Spurious correlations...



Spurious correlations...



!!Pop Quiz!!

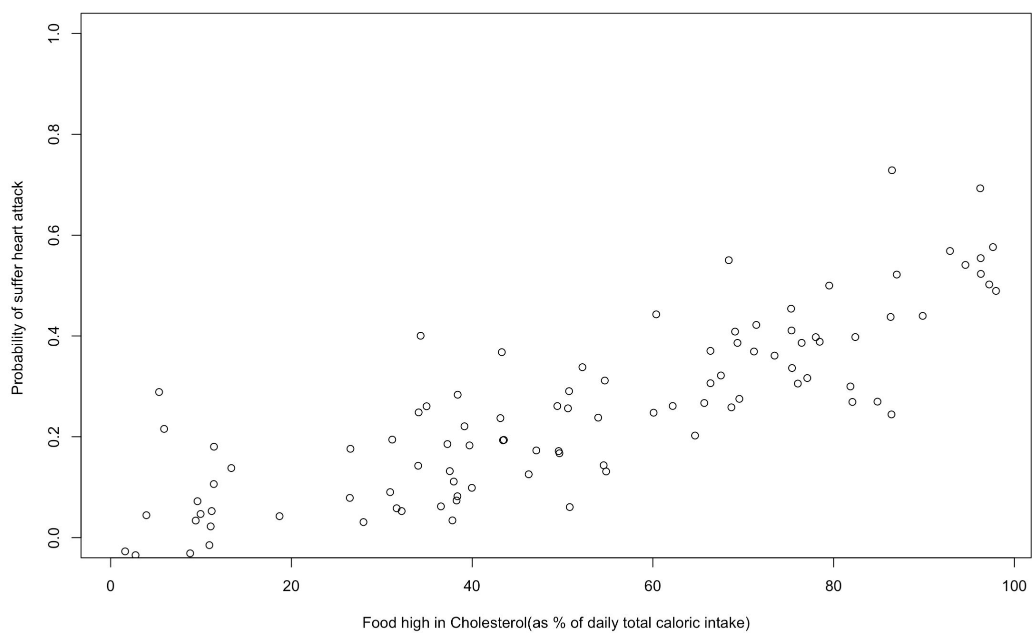
What do you think about the label “spurious correlations”?

What are some possible relationships between association and causation?

Case 1: Association = Causation

Case 1 (Direct Causation) X → Y

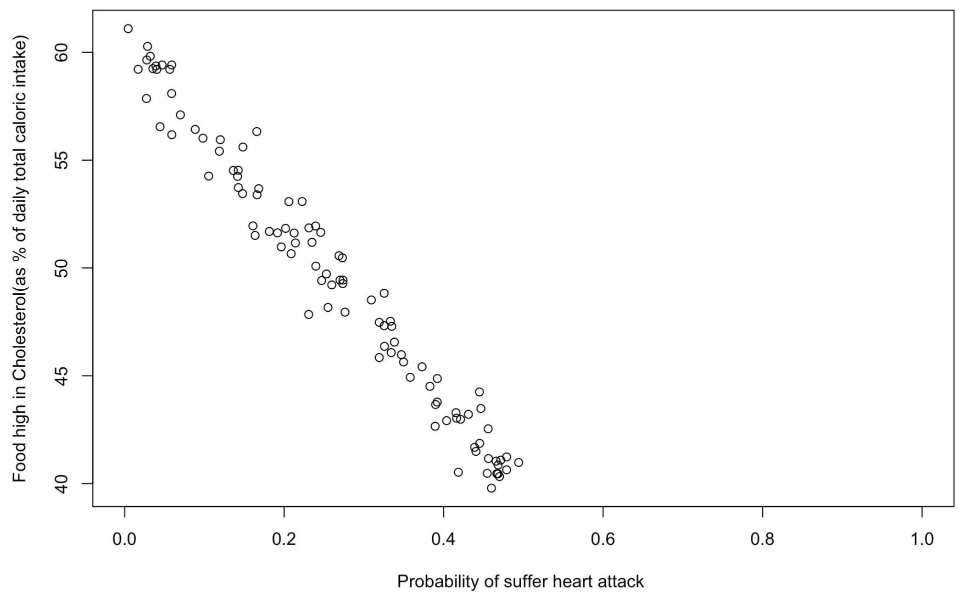
By increasing the intake of foods high in cholesterol (X), people will have higher risk of having a heart attack (Y) as compared to a scenario in which they did not increase the intake of foods high in saturated cholesterol etc.



Case 2: Negative association = positive effect

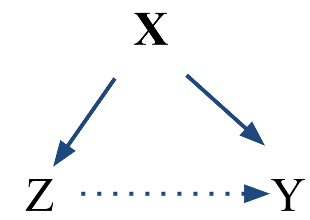
Case 2 (Reverse Causation): X ← Y

Now, suppose when people become concerned about their high risk of having a heart attack (Y) they are more likely than the general population to reduce their intake of foods high in saturated fat and cholesterol (X). You might see a negative association even though there is a positive effect!



Case 3: Negative/positive association = no effect

Case 3 (Confounder) :



If *X* predicts both Z and *Y*, Z and Y can be associated

without being causally related at all.

The presence of a common "cause" is often called *confounding*.

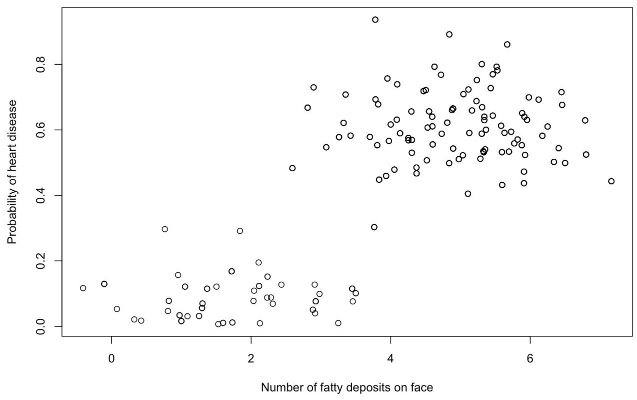
X here is called a confounder.

Following our previous example,let X be a dichotomous variable that denotes whether a patient has high cholesterol intake, all other body condition are identical. Let Z be the number of fatty deposits on face , Let Y be the probability of heart disease. In this simple example, conditioning on the levels of X will eliminate the association between Z and Y.

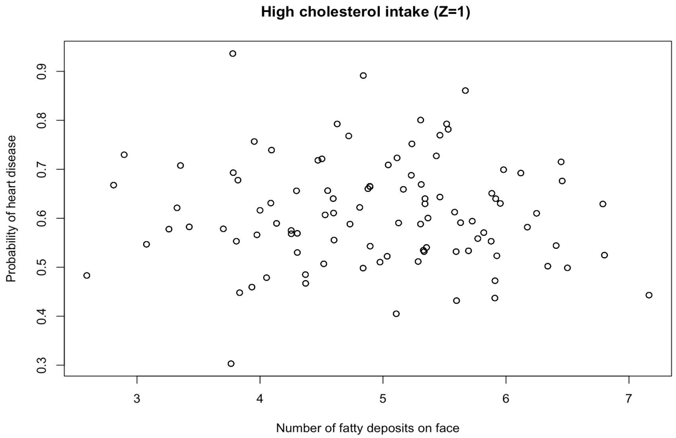
Challenge

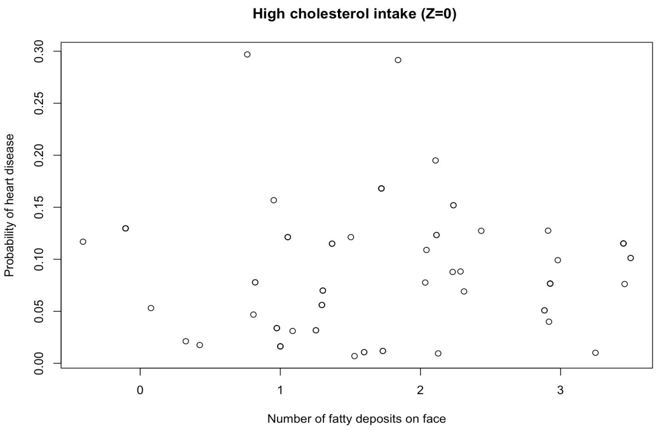
Can you draw a scatterplot of data that has the properties described in the previous slide? (Hint: you can treat Z as a binary variable)

How can two variables be associated in the population?



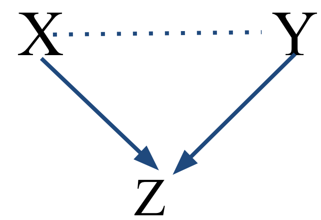
Fatty deposits cause heart disease ?
Their correlation is about 0.5209
What's wrong ?
Let's stratify the data based on level of cholesterol intake (Z=0 and Z=1)
When Z=1, correlation becomes -0.04.
When Z=0, correlation is 0.023.





How can two variables be associated in the population?

Case 4 (Colliders)



A variable is called a collider when

it is causally influenced by two or more variables (both X and Y are causing Z)

- Let's follow our previous example and assume there are only two cause of heart disease congenital heart defects and high level of cholesterol intake.
- X is the level of congenital heart defects .
- Y is level of cholesterol intake as percentage of total caloric intake.
- Z is probability of heart disease.
- There is obviously no correlation between X and Y. But what would happened if condition on level of Z ? For example , I note that people have high probability of heart disease .Then , I know that if it is did not congenital heart defects , that people must have high level of cholesterol intake . X and Y are now no longer independent !

If you see that Z and Y are associated what could that mean?

Two variables Z and Y will be associated in the population if

- Z causes Y. (Case 1)
- Y causes Z.(Case 2)
- There is a X that is a cause of both Z and Y. (Case 3: Confounders)
- (Other special cases)

Take away message:

“association does not imply causation” !

Why do we care about causal inference?

Most research questions are causal questions

What Drives Success?

Does exposing preschoolers to music make them smarter?

Can we alter genes to repel HIV?

Is obesity contagious?

Grief Can Cause a Heart Attack

Did the introduction of CitiBike make New Yorkers healthier?

Does the death penalty reduce crime?

What Happens When the Poor Receive a Stipend?

The New York Times

Hubris

WIRED MAGAZINE: 16.07

Science : Discoveries

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

The Petabyte Age:

"There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show."

Why is Causal Inference Important in Policy?

- Policy decision makers rely on evidence to design policy intervention
- Policies must reflect an accurate causal relationship because they will affect health, education, environmental and other outcomes for millions
- Otherwise, it can lead to loss of resources and lives. Examples:
  - Online ads and purchases
  - Hormone Replacement Therapy: Nurses Health Study versus Women’s Health Initiative

Cautionary Tale: Search Engine Marketing

- $31.7 billion was spent in the U.S. in 2011 on internet advertising
- Common wisdom in the marketing world is that internet advertising is highly effective (can position ads based on type of browsing activity to better target interested shoppers)
- Impact of this type of advertising has been considered easier to measure because can track info on those who click on ads (including "Did they buy?")
- Prediction models suggest that clicking on the ads strongly increase probability of success (reaching website, buying product)
- However what is difficult to measure by just observing the (unmanipulated) system is whether shoppers would have reached the relevant website or bought the product *anyway*

From Blake, T., Nosko, C., and S. Tadelis (2013) "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment"

Cautionary Tale: SEM (cont)

- Researchers now have conducted quasi-experiments to investigate
- Question: What are the returns of brand keyword search advertising? (e.g. "ebay shoes")
- Study design: In March 2012 eBay
  - halted advertising of its brand related terms on Yahoo! and MSN, but
  - continued advertising on Google.
- Results: 99.5% of click traffic was simply redirected through "natural" (unpaid) search traffic. i.e. almost everyone found the site anyway
- Results from other studies showed that non-brand keyword ads have a positive effect on new/infrequent users but no effect for the more frequent users (who represent most of the cost), thus average returns are *negative*.
- From Blake, T., Nosko, C., and S. Tadelis (2013) "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment"

Hormone Replacement Therapy

- Nurses Health Study (massive long-term observational study) shows benefits of HRT for coronary heart disease
- Women's Health Initiative (randomized experiment) shows the opposite results
- Hernan and Robins (*Epidemiology*, 2008) use thoughtful statistical analyses and careful causal thinking to reconcile results (a win for statistical causal inference!)

Example: effect of salt on hypertension

Some city governments have enacted policy to help prevent the detrimental effects of excess salt on health. What is the evidence for this connection?

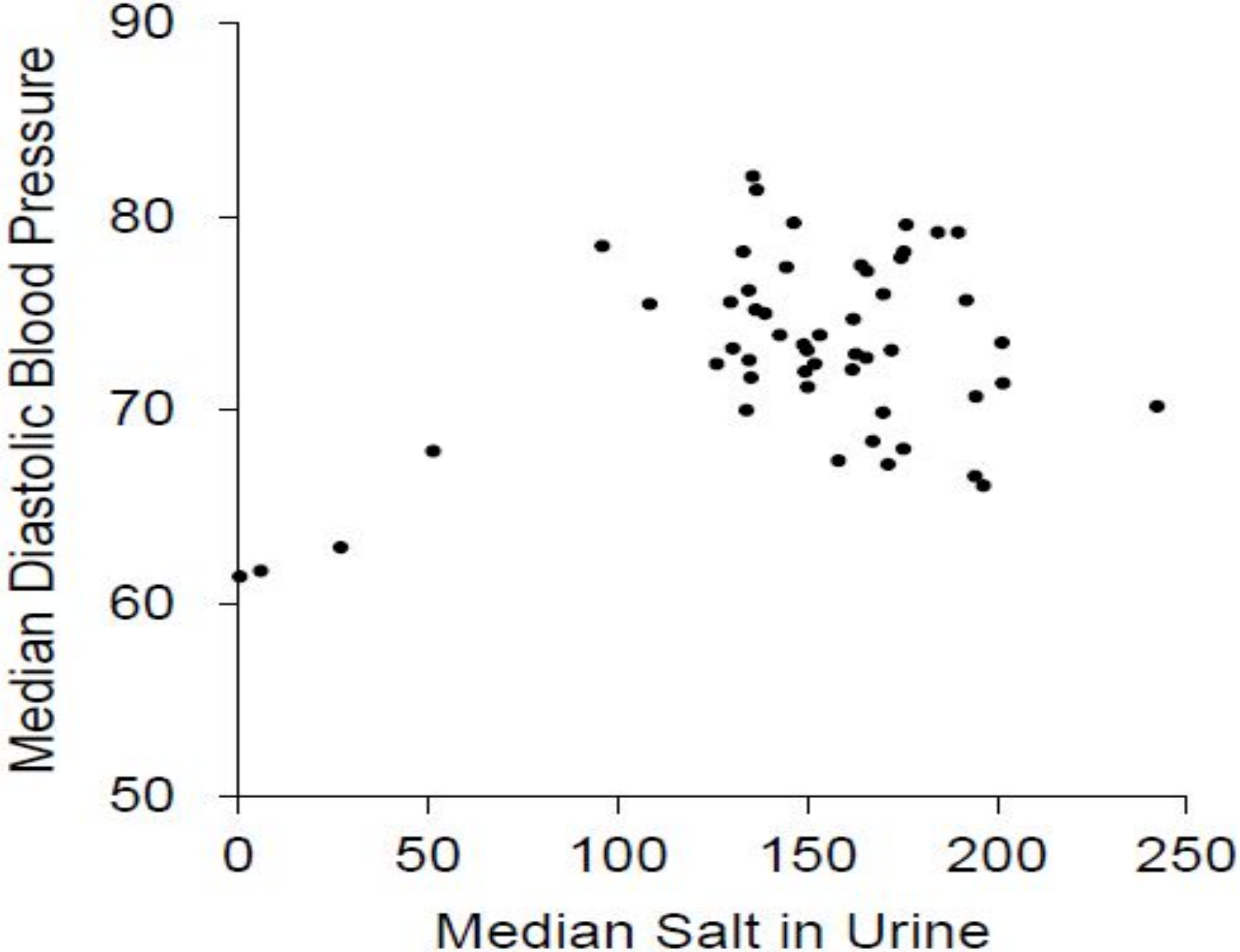
One of the primary sources of evidence used in the Intersalt Trial.

Example: Intersalt Trial

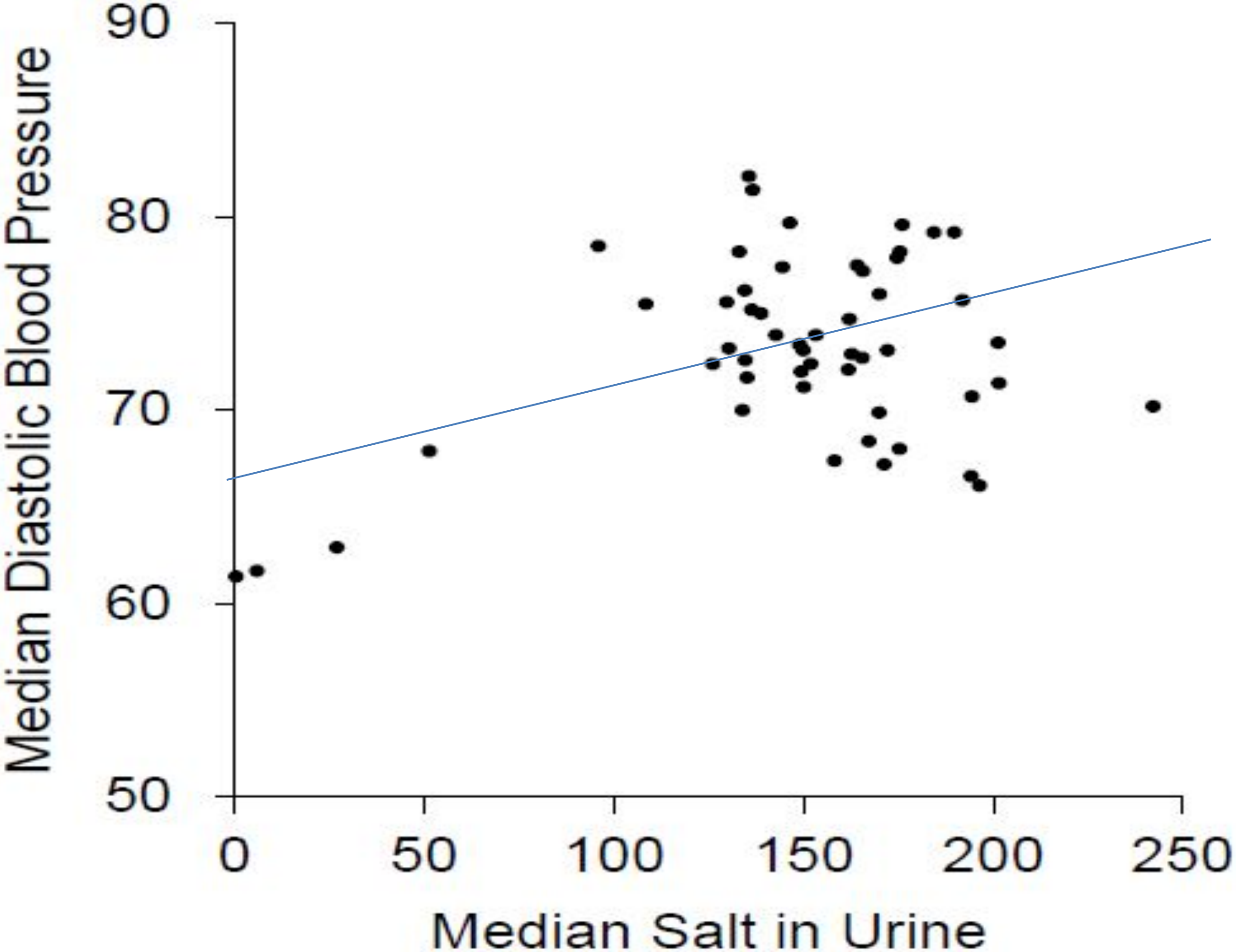
“Intersalt was an observational study, conducted at 52 centers in 32 countries; about 200 subjects age 20–59 were recruited in each center. The two Brazilian centers were Indian tribes, the Yanomamo and Xingu. There was a center in Kenya, and one in Papua New Guinea. In Canada, there were centers in Labrador and in St. John’s (Newfoundland). In the United States, there was a center in Hawaii, a center in Chicago, and four centers in Mississippi. Blood pressure (systolic and diastolic) was measured for each subject, along with urinary sodium and potassium (mmols/24 hrs)…”

(From Freedman et al.)

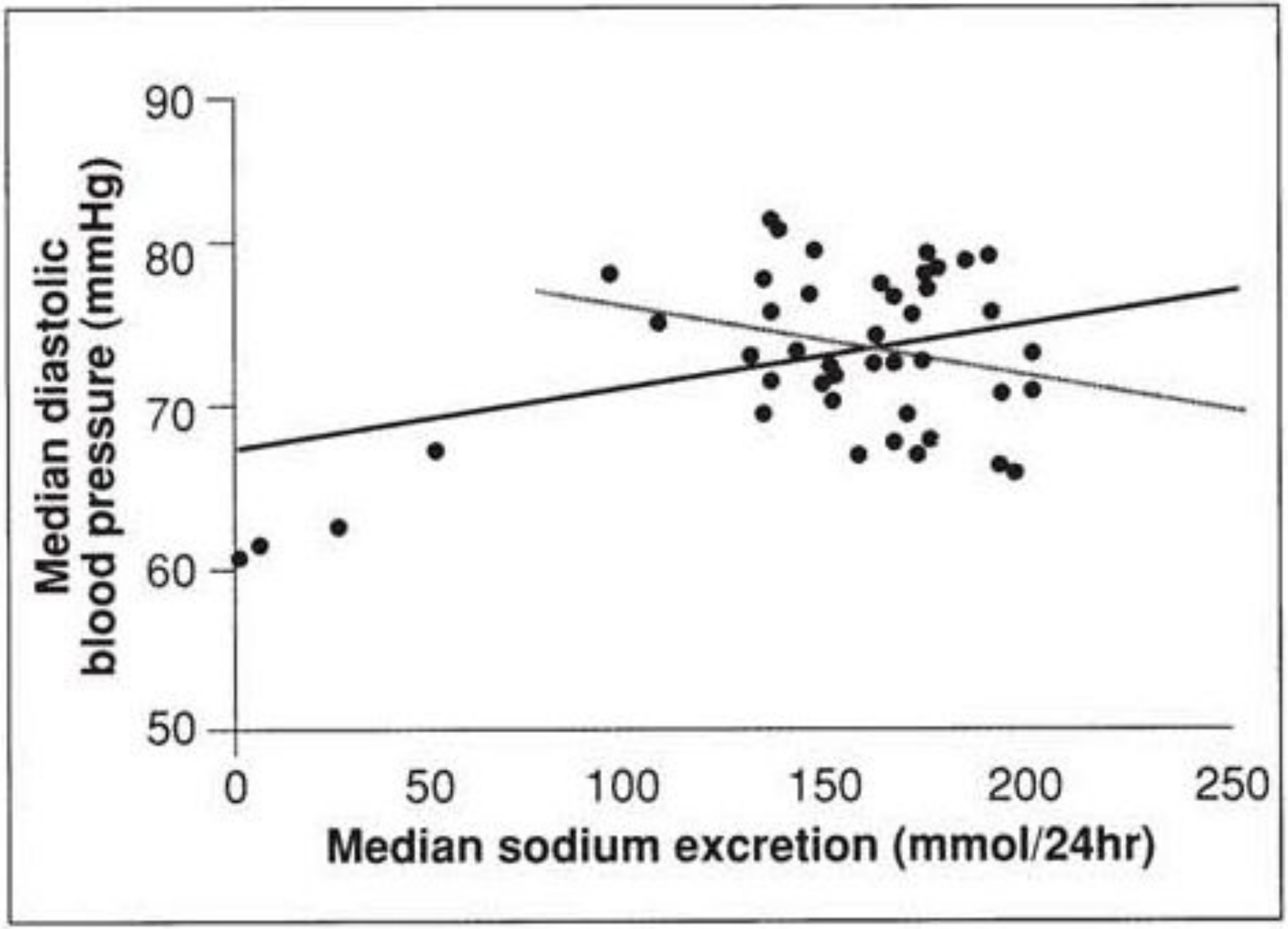
Example: Intersalt Trial



Example: Intersalt Trial



Example: Intersalt Trial



Formalizing Causality Using the Potential Outcomes Framework

How do we define a causal effect?
Example 1:

Suppose you have a headache.

You take some ibuprofen.

An hour later the headache is gone.

Did the ibuprofen cause the headache to go away?

How do we define a causal effect? Example 2:

Research Q: Do fish oil supplements reduce blood pressure?

To answer for person i what would we need to know?

How do we define a causal effect?
Example 2:

Research Q: Do fish oil supplements reduce blood pressure?

To answer for person *i* we would need to know both…..

Potential Outcomes:

$\boldsymbol{Y_i(1)}$: The blood pressure that would result if the person had received the prescribed supplement (denoted $\boldsymbol{y_i}(1)$ in G&H)

$\boldsymbol{Y_i(0)}$: The blood pressure that would result if the person had **not** received the prescribed supplement (denoted $\boldsymbol{y_i}(0)$ in G&H)

So that we could compare them, as in: $\boldsymbol{Y_i(1)-Y_i(0)}$

The Fundamental Problem of Causal Inference,

To estimate a causal effect we need both $Y(0)$ and $Y(1)$

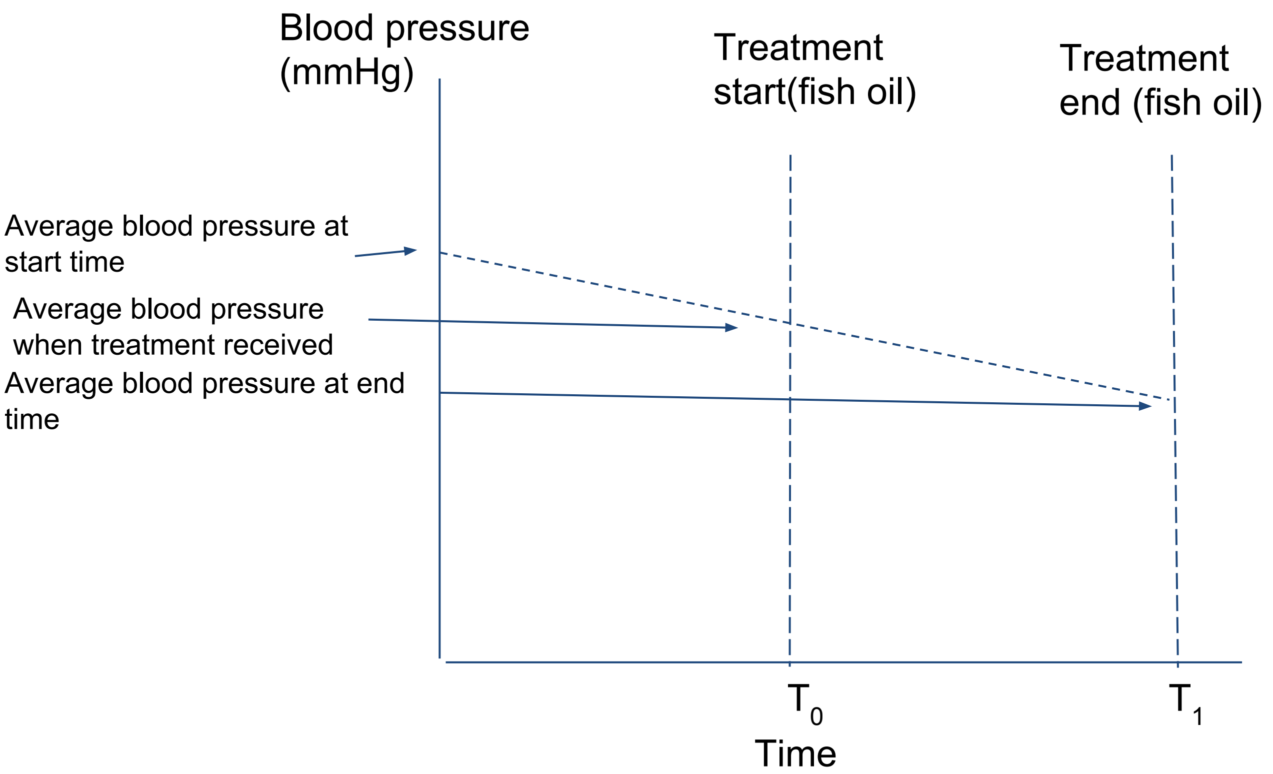
Can you ingest the fish supplement and not ingest the fish supplement at the same time to measure both potential outcomes?

The causal effect is **impossible** to measure directly because we **cannot** observe $Y_i(0)$ and $Y_i(1)$ at the **same time for the same person**

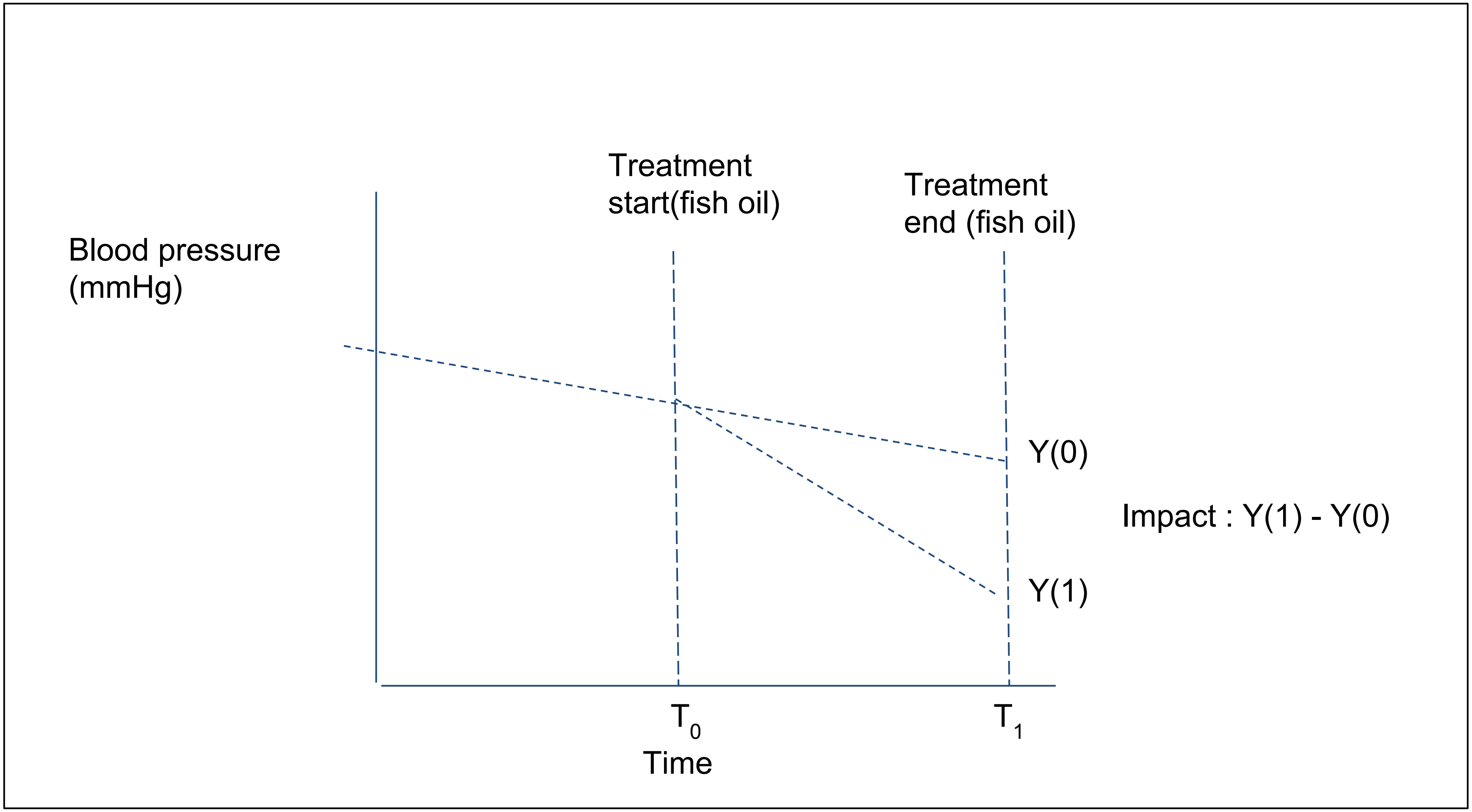
Visualizing counterfactuals

The plot below depicts the trajectory for one person

Can we observe Y(0) or Y(1) in this plot?



Visualizing counterfactuals



Counterfactuals and Potential outcomes

Potential outcomes	Do we observe it?
Y(1) when Z = 1	
Y(1) when Z = 0	
Y(0) when Z = 1	
Y(0) when Z = 0	

Typically this is the type of information a researcher would have:

Unit i	**Female, x_{1i}**	**Age, x_{2i}**	**Treatment, z_i**	[redacted]	[redacted]	**Observed outcome, y_i**
Audrey	1	40	0	[redacted]	[redacted]	140
Anna	1	40	0	[redacted]	[redacted]	140
Bob	0	50	0	[redacted]	[redacted]	150
Bill	0	50	0	[redacted]	[redacted]	150
Caitlin	1	60	1	[redacted]	[redacted]	155
Cara	1	60	1	[redacted]	[redacted]	155
Dave	0	70	1	[redacted]	[redacted]	160
Doug	0	70	1	[redacted]	[redacted]	160

This is what the god of Statistics would see

Unit i	**Female, x_{1i}**	**Age, x_{2i}**	**Treatment, z_i**	**if z_i=0, y^0_i**	**if z_i=1, y_i^1**	**Observed outcome, y_i**
Audrey	1	40	0	**140**	135	140
Anna	1	40	0	**140**	135	140
Bob	0	50	0	**150**	140	150
Bill	0	50	0	**150**	140	150
Caitlin	1	60	1	160	**155**	155
Cara	1	60	1	160	**155**	155
Dave	0	70	1	170	**160**	160
Doug	0	70	1	170	**160**	160

Back to researcher view: Where art thou, Counterfactuals?

Unit i	**Female, x_{1i}**	**Age, x_{2i}**	**Treatment, z_i**	**if z_i=0, y^0_i**	**if z_i=1, y_i^1**	**Observed outcome, y_i**
Audrey	1	40	0	**140**	?	140
Anna	1	40	0	**140**	?	140
Bob	0	50	0	**150**	?	150
Bill	0	50	0	**150**	?	150
Caitlin	1	60	1	?	**155**	155
Cara	1	60	1	?	**155**	155
Dave	0	70	1	?	**160**	160
Doug	0	70	1	?	**160**	160

Q: Is this enough information to fill in the rest?

Unit i	**Female, x_{1i}**	**Age, x_{2i}**	**Treatment, z_i**	**if z_i=0, y^0_i**	**if z_i=1, y_i^1**	**Observed outcome, y_i**
Audrey	1	40	0	**140**	135	140
Anna	1	40	0	**140**	?	140
Bob	0	50	0	**150**	?	150
Bill	0	50	0	**150**	?	150
Caitlin	1	60	1	?	**155**	155
Cara	1	60	1	?	**155**	155
Dave	0	70	1	?	**160**	160
Doug	0	70	1	?	**160**	160

So how do we proceed?

So how do we proceed?

One solution is a randomized experiment.

The goal is to translate the following

Randomized Experiments: God vision

Unit i	**Female, x_{1i}**	**Age, x_{2i}**	**Treatment, z_i**	**if z_i=0, y^0_i**	**if z_i=1, y_i^1**	**Observed outcome, y_i**
Audrey	1	40	0	**140**	135	140
Anna	1	40	1	140	**135**	135
Bob	0	50	0	**150**	140	150
Bill	0	50	1	150	**140**	140
Caitlin	1	60	0	**160**	155	160
Cara	1	60	1	160	**155**	155
Dave	0	70	0	**170**	160	170
Doug	0	70	1	170	**160**	160

into something the researcher can use to get an unbiased estimate of the treatment effect, as in

Randomized Experiments: Researcher vision

Unit i	**Female, x_{1i}**	**Age, x_{2i}**	**Treatment, z_i**	**if z_i=0, y^0_i**	**if z_i=1, y_i^1**	**Observed outcome, y_i**
Audrey	1	40	0	**140**		140
Anna	1	40	1		**135**	135
Bob	0	50	0	**150**		150
Bill	0	50	1		**140**	140
Caitlin	1	60	0	**160**		160
Cara	1	60	1		**155**	155
Dave	0	70	0	**170**		170
Doug	0	70	1		**160**	160

How are potential outcomes related to the rest of the semester?

Basically every method we are going to talk about can be conceived up as a way of filling in those potential outcomes!

First we'll talk about randomized experiments…..

(wait…. but first)
This Course

Causal study designs and analysis approaches we will study this semester

Randomized experiments

Observational studies with covariates

- stratification
- regression
- matching

Instrumental variables

Regression discontinuity

Difference in differences

Fixed Effects

Sensitivity Analysis

[Bayesian nonparametric approaches (machine learning)]

Course Motivation

- The goal of program evaluation and policy research is to determine the effects of a program or policy already in place, or one that might be implemented, for a given population
- This can be difficult to determine in the absence of a randomized experiment because we generally can't observe what *would have* happened to the population of interest had they been exposed to a different policy
- Even when using randomized experiments these questions can be difficult to answer (missing data, noncompliance, lack of generalizability,…)

Course Goals

- Understand the meaning of a causal estimand
- Understand the theoretical advantages of a randomized experiment
- Understand the practical limitations of a randomized experiment
- Understand the assumptions required for a variety of non-experimental methods to produce causal estimands
- Be able to implement and diagnose a variety of current methods that attempt to estimate causal effects in observational studies

Course style and substance

This course may represent a departure from what you may have been taught in the past both in terms of style and substance

- *Style*. This is a class for motivated students. There is no text book. You will need to learn from our lectures and from the assigned journal articles and book chapters which may not be written at a basic level. I will attempt to provide all the tools you will need but you will also have to show initiative.
- *Substance*. I will attempt to convince you to discard some of the statistical techniques you've become familiar and comfortable with in favor of methods that are better geared for causal inference. Sometimes I may try to convince you that a causal question simply can't be addressed by the data. You will be forced to look at research questions more deeply and thoughtfully. This can be a painful process! But I think it is worth the journey.

Homework Option 1 - Stata track

Assignment 1: (done)

Assignments 2 and 3 reinforce new concepts (potential outcomes, average treatment effects, randomized experiments, simple estimators) through online quizzes that involve algebra and questions about basic concepts.

Assignments 4-7 address propensity scores, instrumental variables, regression discontinuity, fixed effects and difference in differences; these will involve both data analyses (either Stata or R can be used) and a thoughtful description of the assumptions and findings.

Final: Students can choose between a take-home final exam and a final project

This track is best for students who will continue to use Stata and are focused mostly on how to apply methods to data and correctly write up the results.

Homework Option 2 -- R track

Assignment 1: (done)

Assignments 2-7: The next six assignments will address the same methods as in Stata track but will require simulating data and analyzing it in R. These will help them gain a deeper understanding of the structural and parametric assumption required for each design/method as well as the properties of the corresponding estimators. The first assignment will be quite scaffolded but assignments will increase in level of difficulty over time. (Assignment 2 is broken into 2 parts)

Final: Students will have a choice between an in-class final exam using R and a final project

This track is best for students who want a deeper understanding of the assumptions and the statistical properties of the methods; it will also provide stronger preparation for developing new methods and evaluating the comparative efficacy of existing methods.

Homework Option 2 -- R track-- more information

Assignment 2 for the R track will be broken into 2 parts so that students who are “shopping” the options can try it out before committing.

Assignment 2a will be released today and due next Friday. Assignment 2b will be released soon and due Friday September

Class discussion boards

Please sign up on Piazza to access the class discussion boards!

piazza.com/nyu/fall2018/apstage2012