Multivariate thinking and the introductory statistics and data science course: preparing students to make sense of observational data

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

JSM, July 31, 2018

nhorton@amherst.edu
https://www.github.com/Amherst-Statistics/JSM2018

Thanks and acknowledgements

- Project MOSAIC: Danny Kaplan (Macalester College), Randy Pruim (Calvin College), Ben Baumer (Smith College), and Johanna Hardin (Pomona College)
- Hill, Pearl, Scheines, Robins, Hernan, Rubin, Vanderweele, Winship, Morgan, Tchetgen Tchetgen, Cobb, and many others
- Sarah Anoke, Sonia Hernandez-Diaz, Miguel Hernan, Murray Mittelman, Brendan Seto, and Sonja Swanson for many useful suggestions, curricular models, and examples

Motivation

- Are we teaching what we should be teaching in our introductory statistics and data science courses?

Motivation

- Are we teaching what we should be teaching in our introductory statistics and data science courses?
- OR do students leave our courses and programs with a form of observational data paralysis?
- Are we hiding the power of statistics behind our bushel basket?

Source: causal inference FB

- What evidence do we have that smoking causes lung cancer?

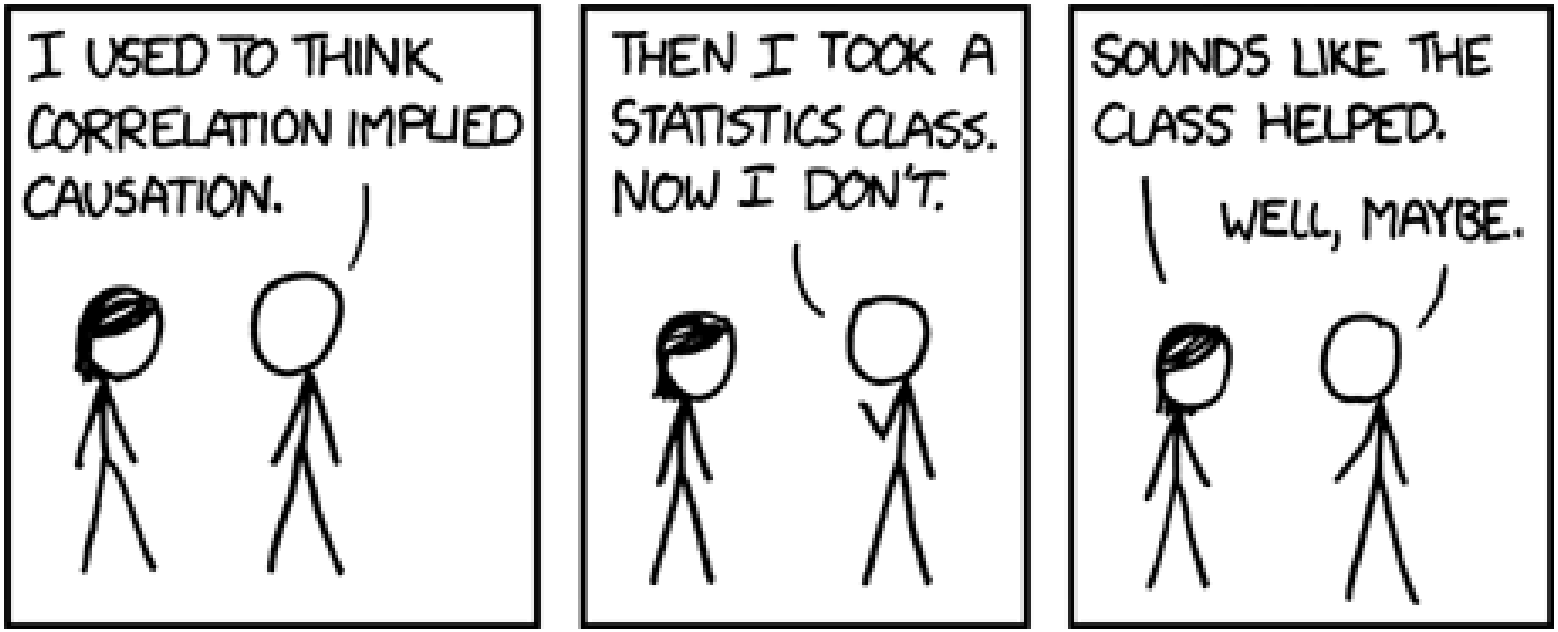
Source: causal inference FB

- While clinical trials are wonderful, we live in a world of 'found' data.
- "It is not that i believe an experiment is the only proper setting for discussing causality, but I do feel that it is the simplest such setting" - Holland (1986)

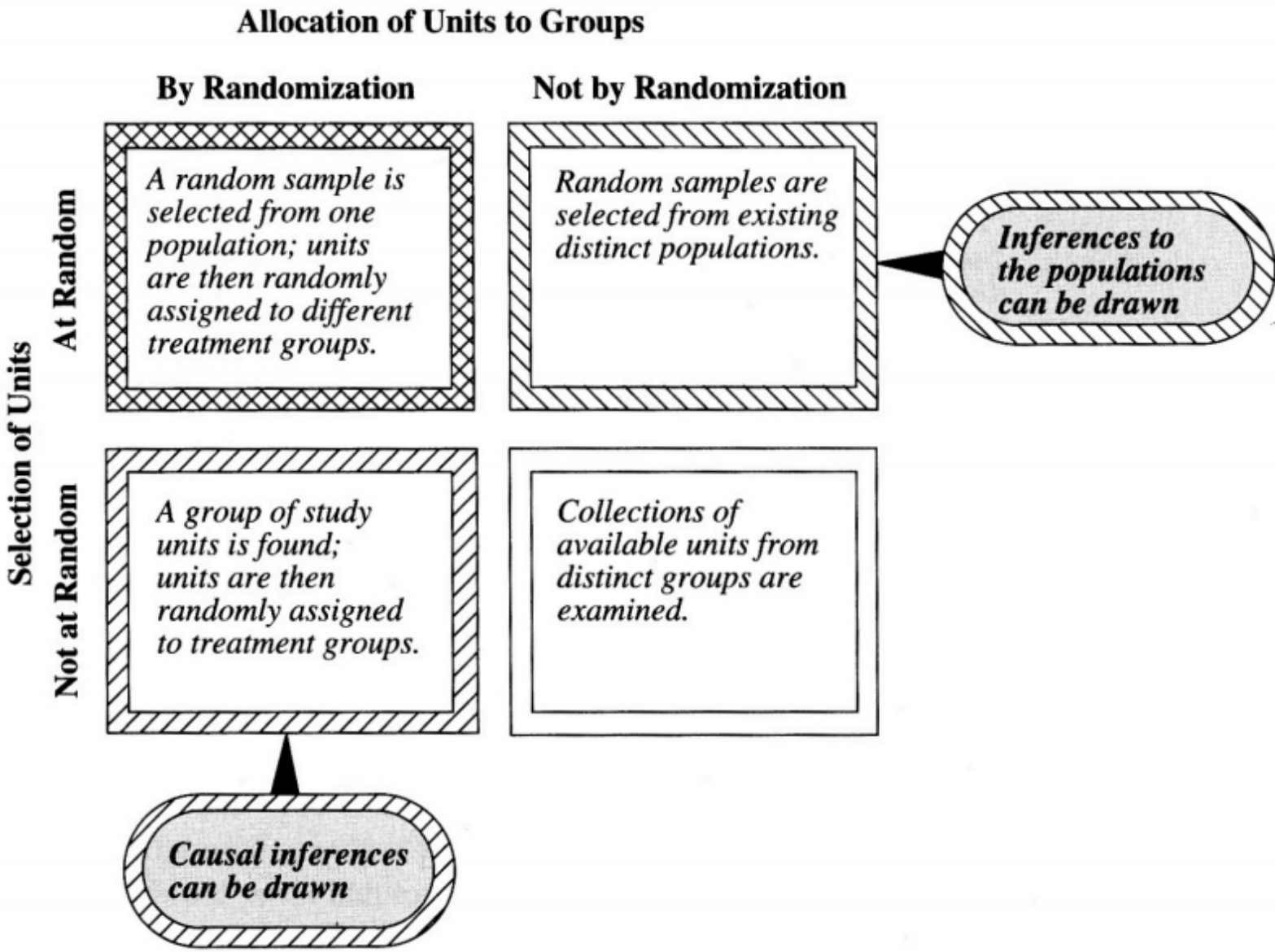
Teaching multivariate thinking and confounding

1. what are we currently teaching?

Do we teach in a way that encourages paralysis?



Other factors may be responsible for observed associations



Example from SDM4 (De Veaux, Velleman, and Bock) p. 575

Exercise 20.41: It's widely believed that regular mammogram screening may detect breast cancer early, resulting in fewer deaths from that disease. One study that investigated this issue over a period of 18 years was published during the 1970's. Among 30,565 who had never had mammograms, 196 died of breast cancer (0.64%) while only 153 of 30,131 who had undergone screening died of breast cancer (0.50%).

Do these results suggest that mammograms may be an effective screening tool to reduce breast cancer deaths?

Solution to Exercise 20.41 SDM4 (De Veaux, Velleman, and Bock) p. 575

$H_0 : p_1 - p_2 = 0$ vs. $H_A : p_1 - p_2 > 0$ (one-sided test? That's a different sermon.)

Solution to Exercise 20.41 SDM4 (De Veaux, Velleman, and Bock) p. 575

$H_0 : p_1 - p_2 = 0$ vs. $H_A : p_1 - p_2 > 0$ (one-sided test? That's a different sermon.) where p_1 is the proportion of women who never had mammograms who died of breast cancer and p_2 is the proportion of women who had undergone screening who died of breast cancer (z=2.17, p=0.0148).

With a p-value this low, we reject H_0. The data suggest that mammograms may reduce breast cancer deaths.

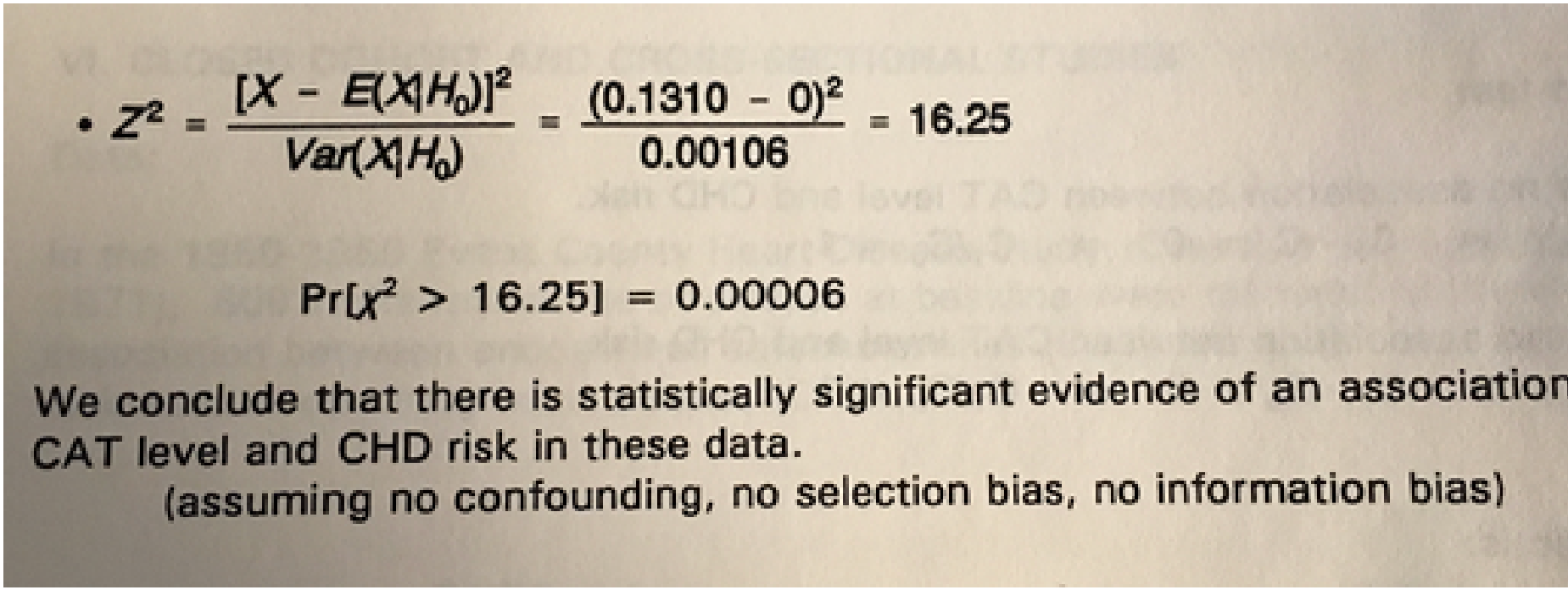
Solution to Exercise 20.41 SDM4 (De Veaux, Velleman, and Bock) p. 575

$H_0 : p_1 - p_2 = 0$ vs. $H_A : p_1 - p_2 > 0$ (one-sided test? That's a different sermon.) where p_1 is the proportion of women who never had mammograms who died of breast cancer and p_2 is the proportion of women who had undergone screening who died of breast cancer (z=2.17, p=0.0148).

With a p-value this low, we reject H_0. The data suggest that mammograms may reduce breast cancer deaths.

(But what about possible confounders?)

Harvard University EPI202b (circa 1995)



$$\bullet\ Z^2 = \frac{[X - E(X|H_0)]^2}{Var(X|H_0)} = \frac{(0.1310 - 0)^2}{0.00106} = 16.25$$

$$Pr[\chi^2 > 16.25] = 0.00006$$

We conclude that there is statistically significant evidence of an association
CAT level and CHD risk in these data.
(assuming no confounding, no selection bias, no information bias)

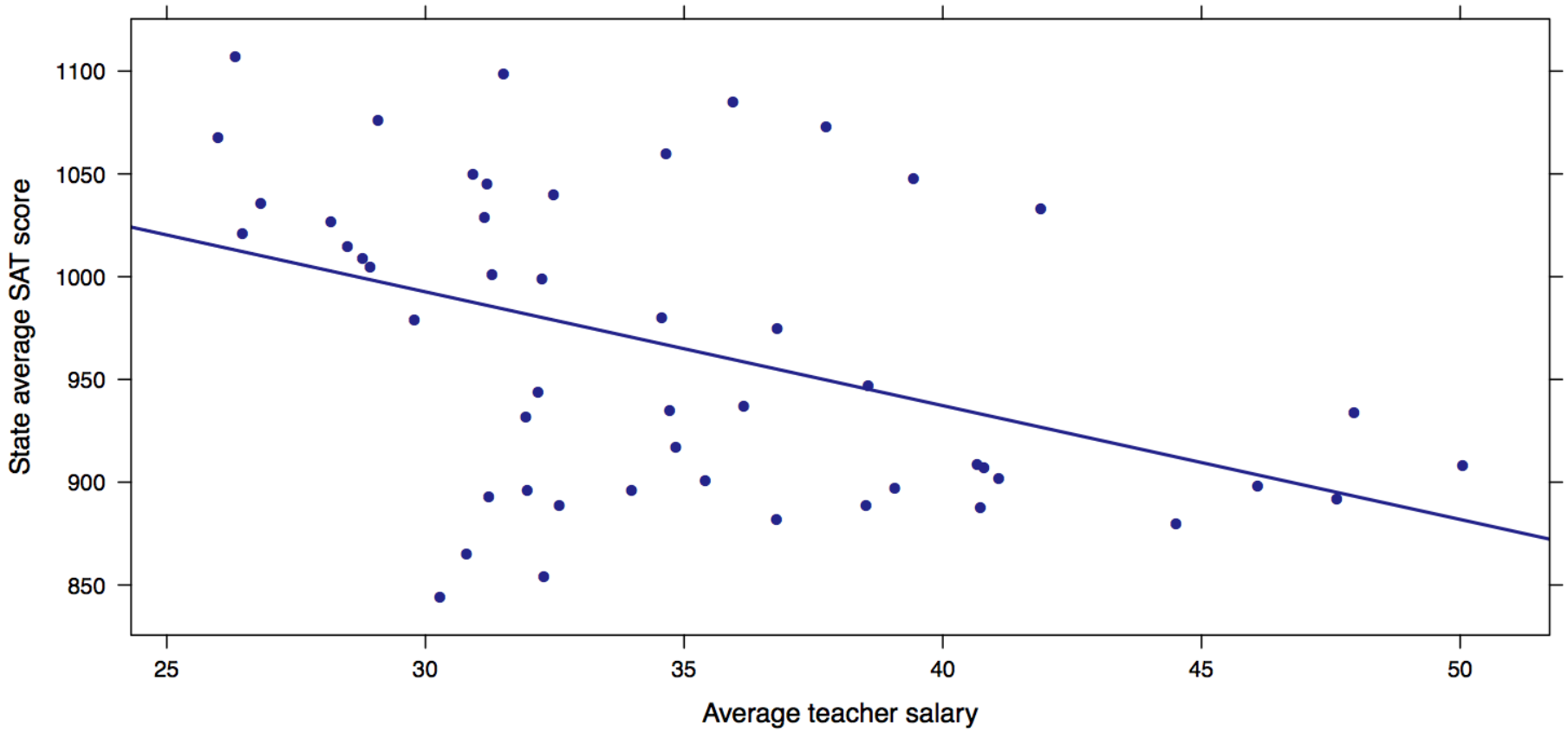
revised GAISE College Report (2016)

1. Teach statistical thinking.
   - Teach statistics as an investigative process of problem-solving and decision-making.
   - Give students experience with **multivariable thinking**.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

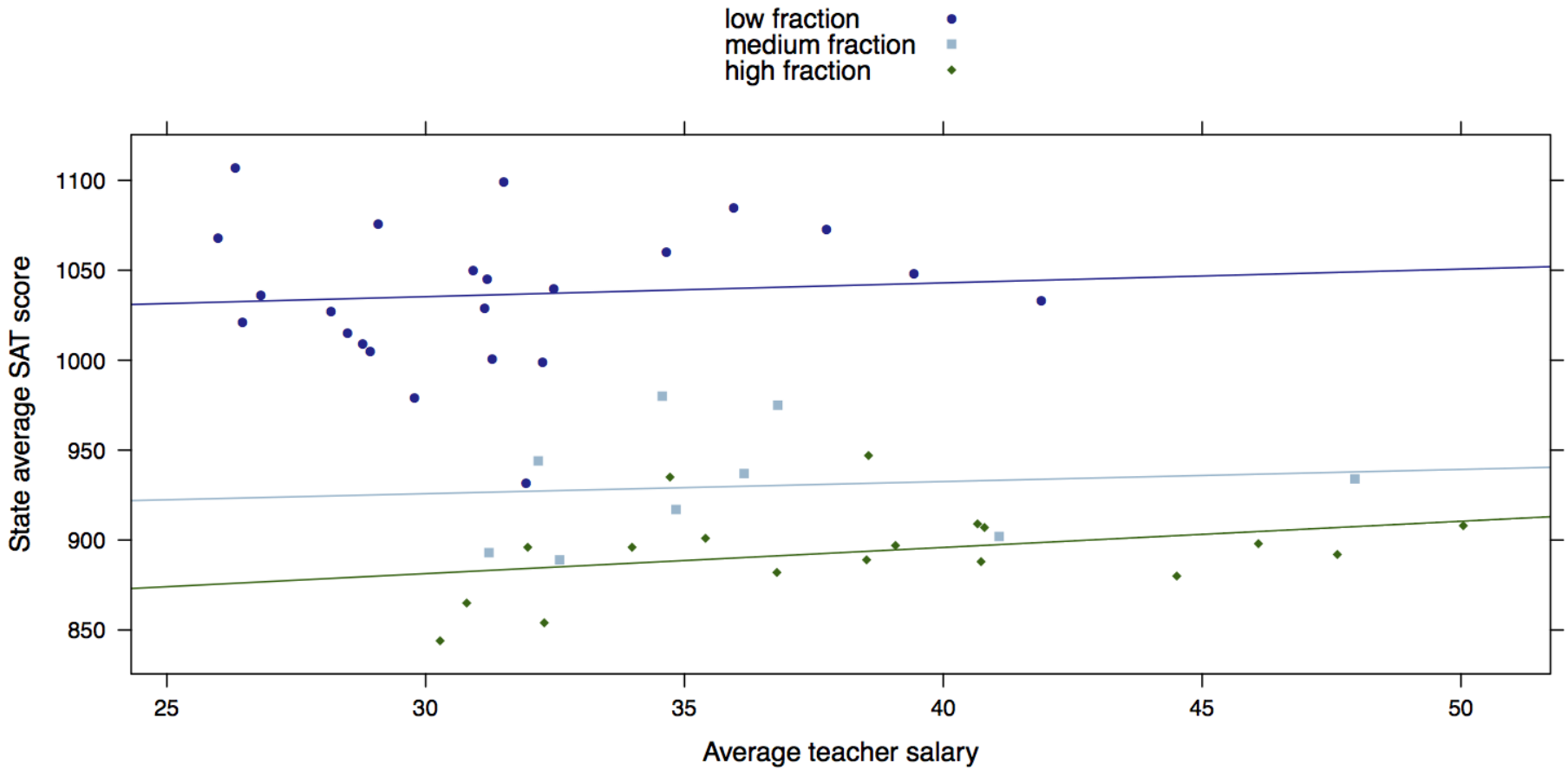
Teaching multivariate thinking and confounding

1. what are we currently teaching?
2. motivating multivariate examples (to build on what Kari Lock Morgan introduced in the prior talk)

SAT scores and teacher salaries (state data from 2010)

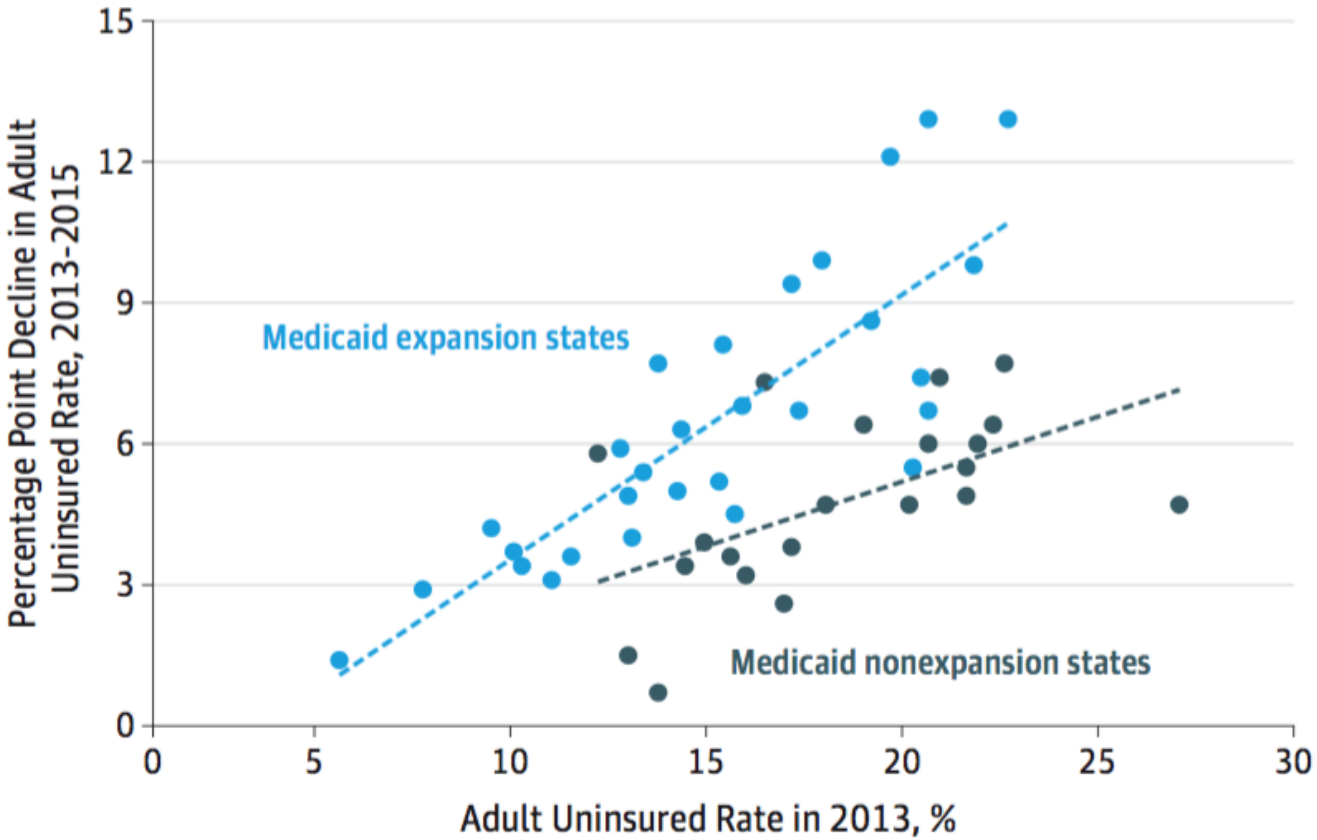


SAT scores and teacher salaries (state data from 2010)



stratification and/or multiple regression: Obama's 2016 single author JAMA paper

Figure 2. Decline in Adult Uninsured Rate From 2013 to 2015 vs 2013 Uninsured Rate by State



See also "Statistical methods in the NEJM" (2007)

How to handle more than two variables?

- "Data Viz on Day One" (TISE, `http://escholarship.org/uc/item/84v3774z`)
- stratification
- multiple regression (early and often)
- straightforward to use mosaic package "Less Volume, More Creativity" approach to modeling (*R Journal*, `https://journal.r-project.org/archive/2017/RJ-2017-024` and related Little Books)
- new emphasis on causal graphs and confounding (more to come on this front)

Teaching multivariate thinking and confounding

1. what are we currently teaching?
2. motivating multivariate examples
3. confounding 101 and 201

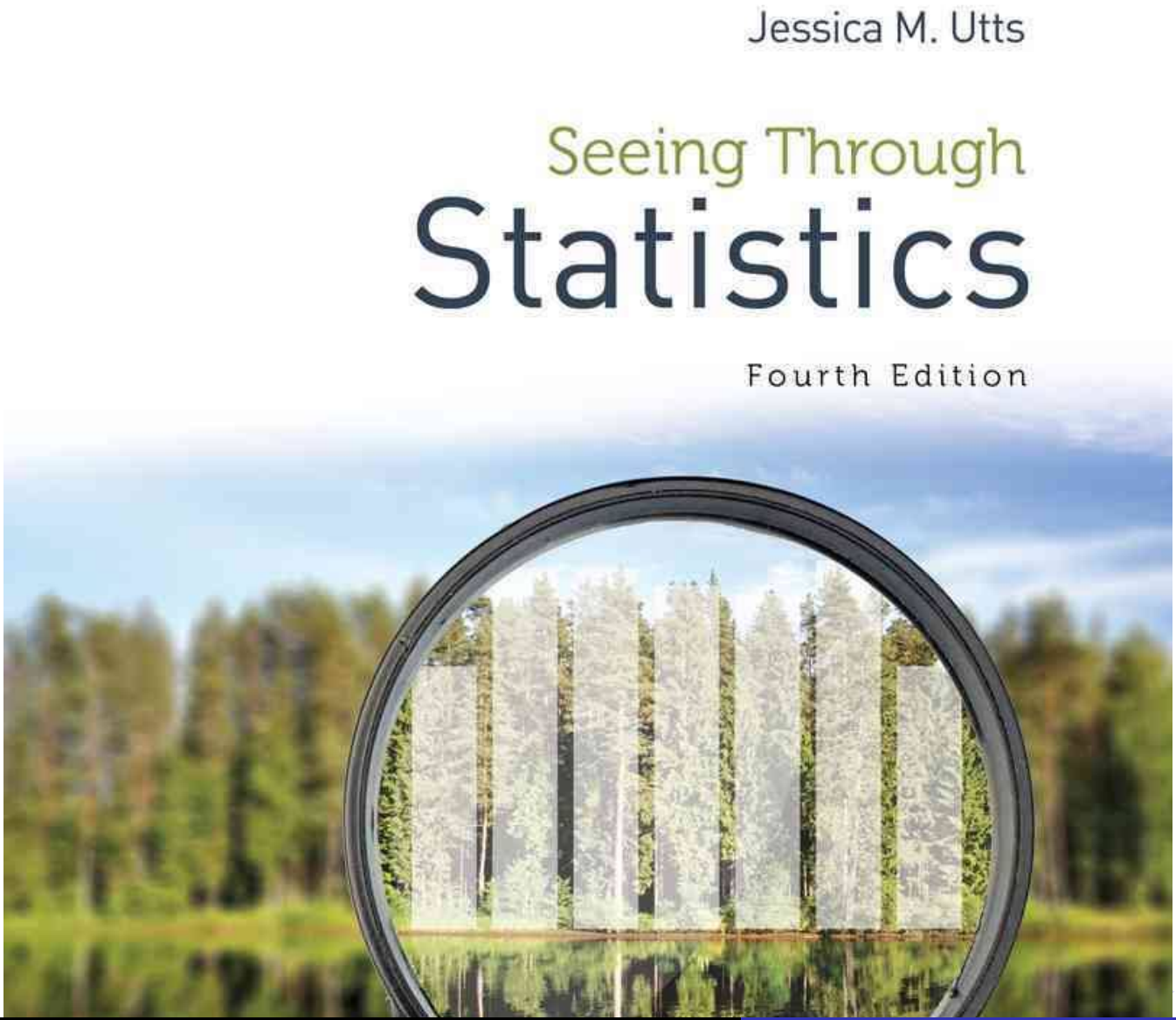
AP Statistics Vocabulary

Both Sides

confounding

when the levels of one factor are associated with the levels of another factor so their effects cannot be separated

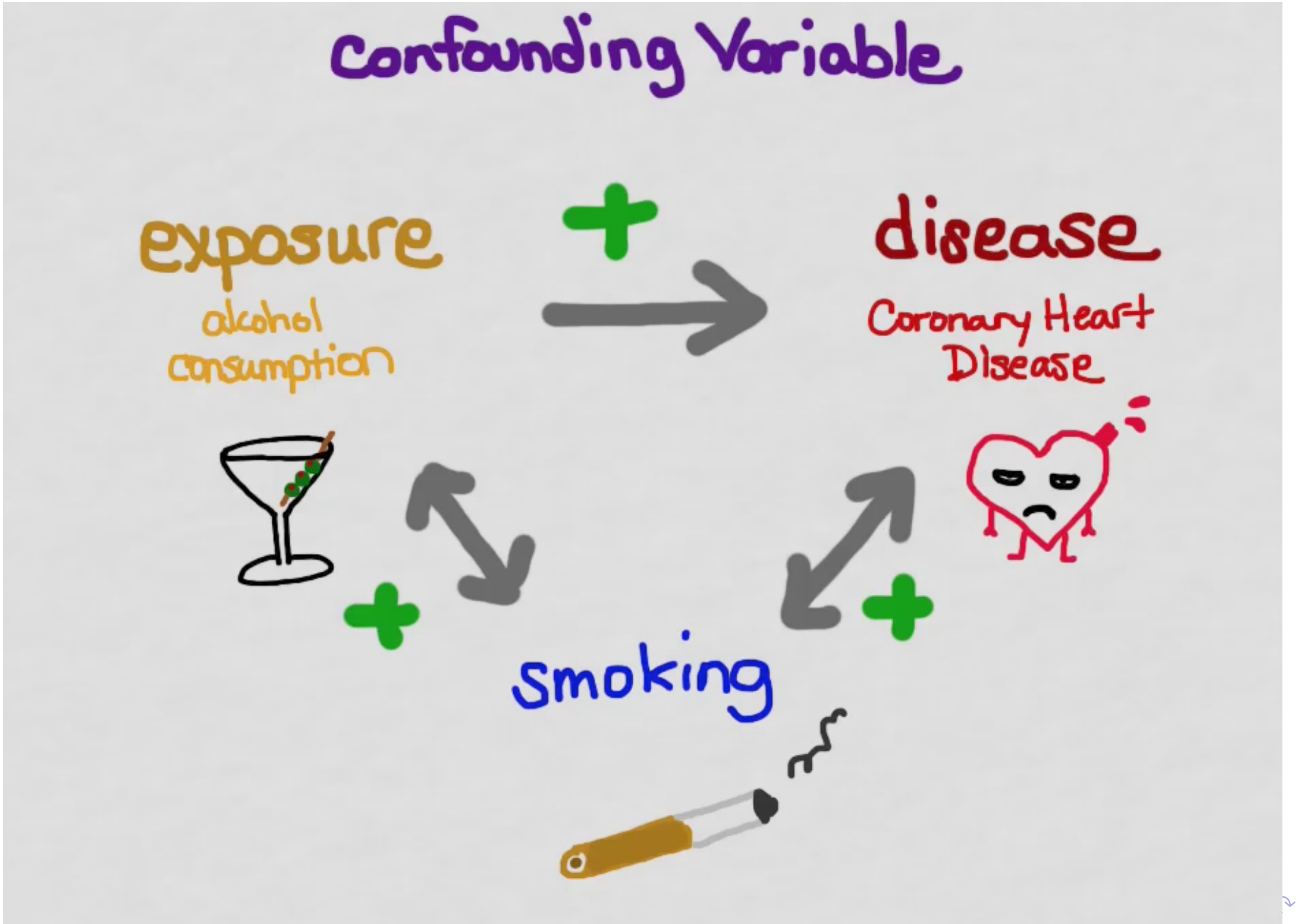
Utts: Seeing through Statistics (4th edition)



A confounding variable is one that has two properties.

1. A confounding variable is related to the explanatory variable in the sense that individuals who differ for the explanatory variable are also likely to differ for the confounding variables.
2. A confounding variable affects the response variable. Because of these two properties, the effect of a confounding variable on the response variable cannot be separated from the effect of the explanatory variable on the response variable.

causal graphs (Elizabeth Lynch)



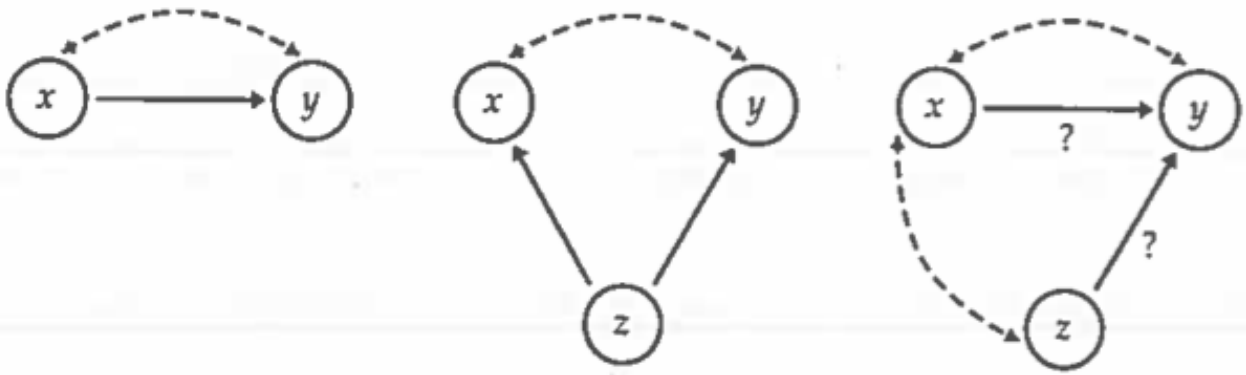
Learning outcomes (Hernan and Swanson, EPI 524)

Confounding is a ubiquitous bias that arises when non-comparable groups are compared. It is one of the greatest threats to valid causal inferences from observational data. Therefore, controlling for confounding is a fundamental component of epidemiologic research.

Explaining association: causation

Figure 2.29 shows in outline form how a variety of underlying links between variables can explain association. The dashed double-arrow line represents an observed association between the variables x and y. Some associations are explained by a direct cause-and-effect link between these variables. The first diagram in Figure 2.28 shows "x causes y" by a solid arrow running from x to y.

Items 1 and 2 in Example 2.42 are examples of direct causation. *Even when direct causation is present, very often it is not a complete explanation of an association between two variables.* The best evidence for causation comes from experiments that actually change x while holding all other factors fixed. If y changes, we have good reason to think that x caused the change in y.

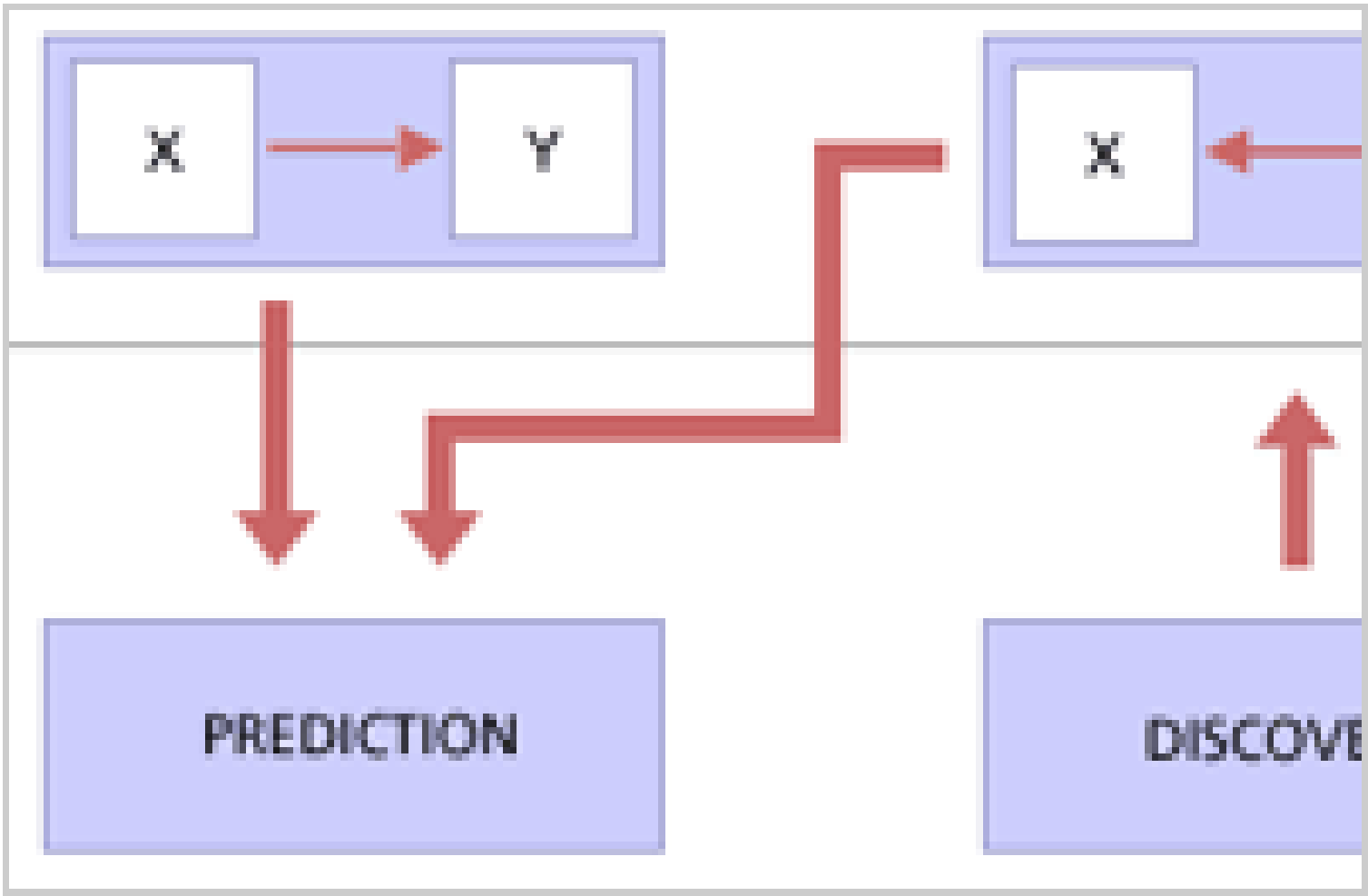


Causation (a) | Common response (b) | Confounding (c)

What to teach?

We need to go beyond these informal definitions...

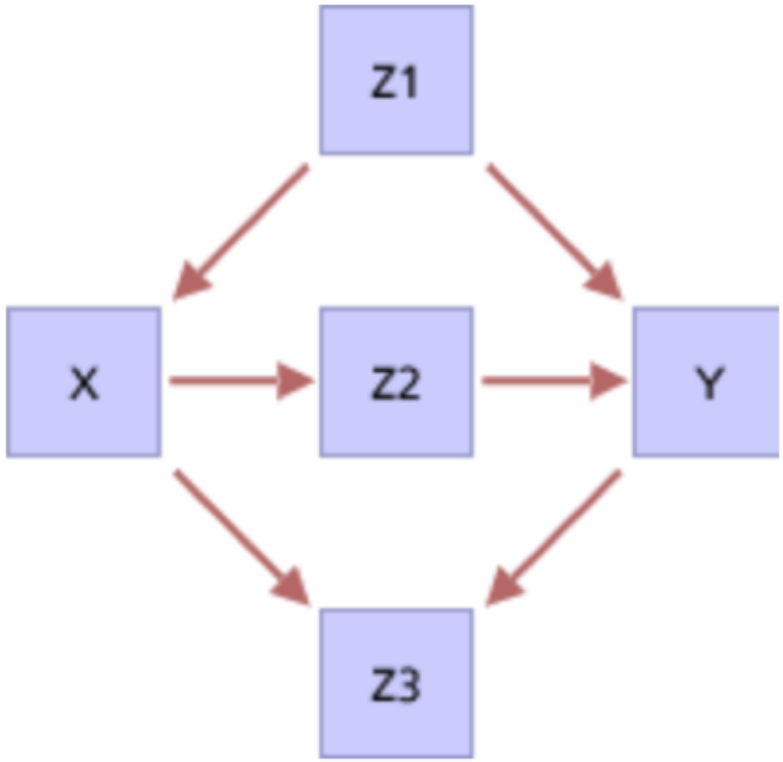
CMU Open Learning Initiative (OLI): Causal and Statistical Reasoning



CMU OLI: Causal and Statistical Reasoning criteria

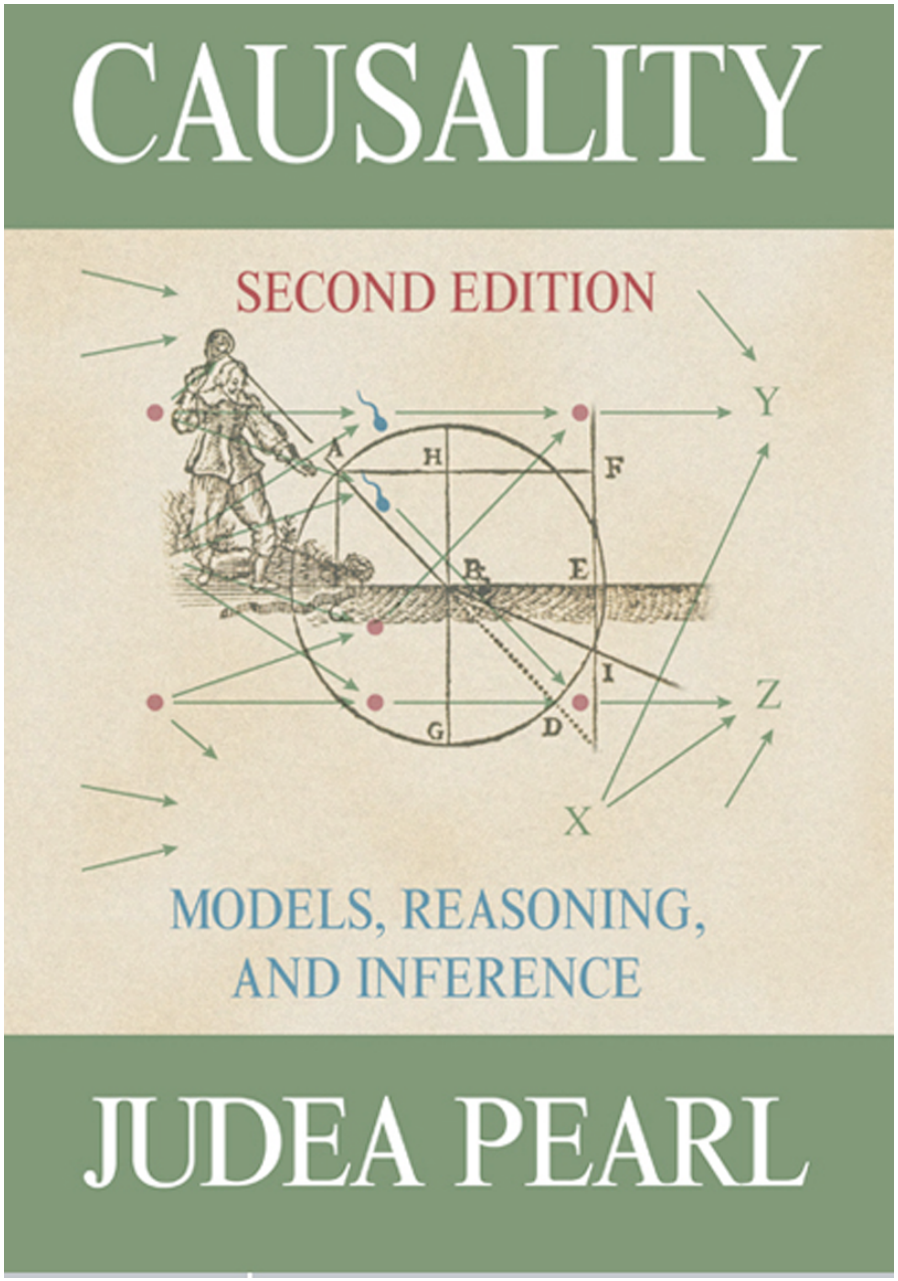
- X and Y are D-separated by Z just in case there are no undirected paths between X and Y that are active relative to Z
- A path is active iff all the variable on the path are active
- Non-colliders are active if they are not in the conditioning set Z, and inactive if they are in Z
- Colliders are active if they are in Z or have an effect in Z, and inactive otherwise

- First identify all undirected paths
- Count number of active (causally connected) paths
  - No mediators or common causes in Z
  - All common effects in Z
- If an active path exists, it is D-connected by that path
- If not, D separated

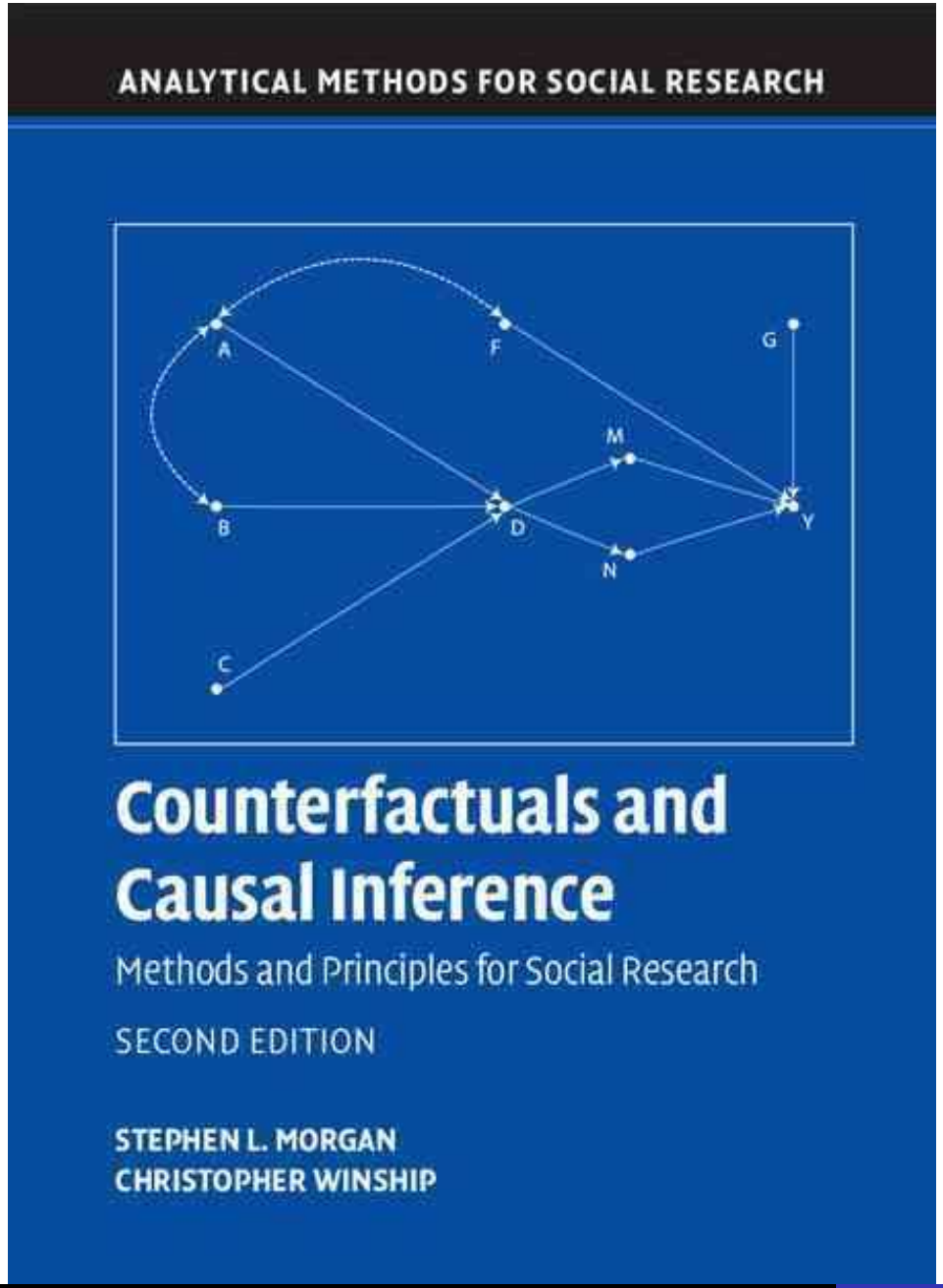


Used these materials in a second course in statistics circa 2009 (but tough going)

Pearl, second edition (2009)



Winship and Morgan, second edition (2015)



6.6 The structure of effect modificati

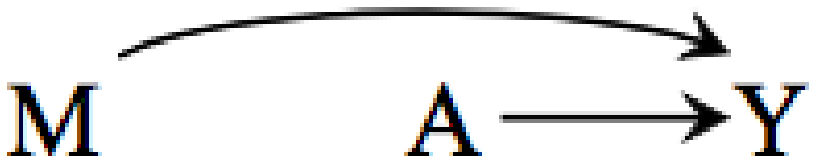


Figure 6.11

Identifying pot
use our causal
association bet
to illustrate th

Suppose he
identify the ave
that there is n
Computing the
association is
$\Pr[Y = 1|A =$

7.1 The structure of confounding

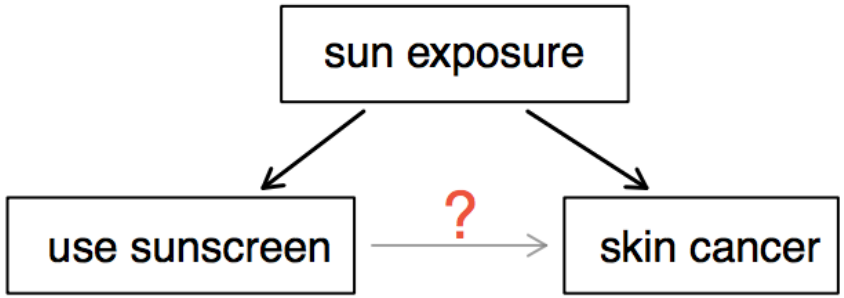
Figure 7.1

Confounding i
a cause. The
diagrams. For
a treatment A
diagram show
the path $A \rightarrow$
$A \leftarrow L \rightarrow Y$
graph theory,
cause L is an

Example

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

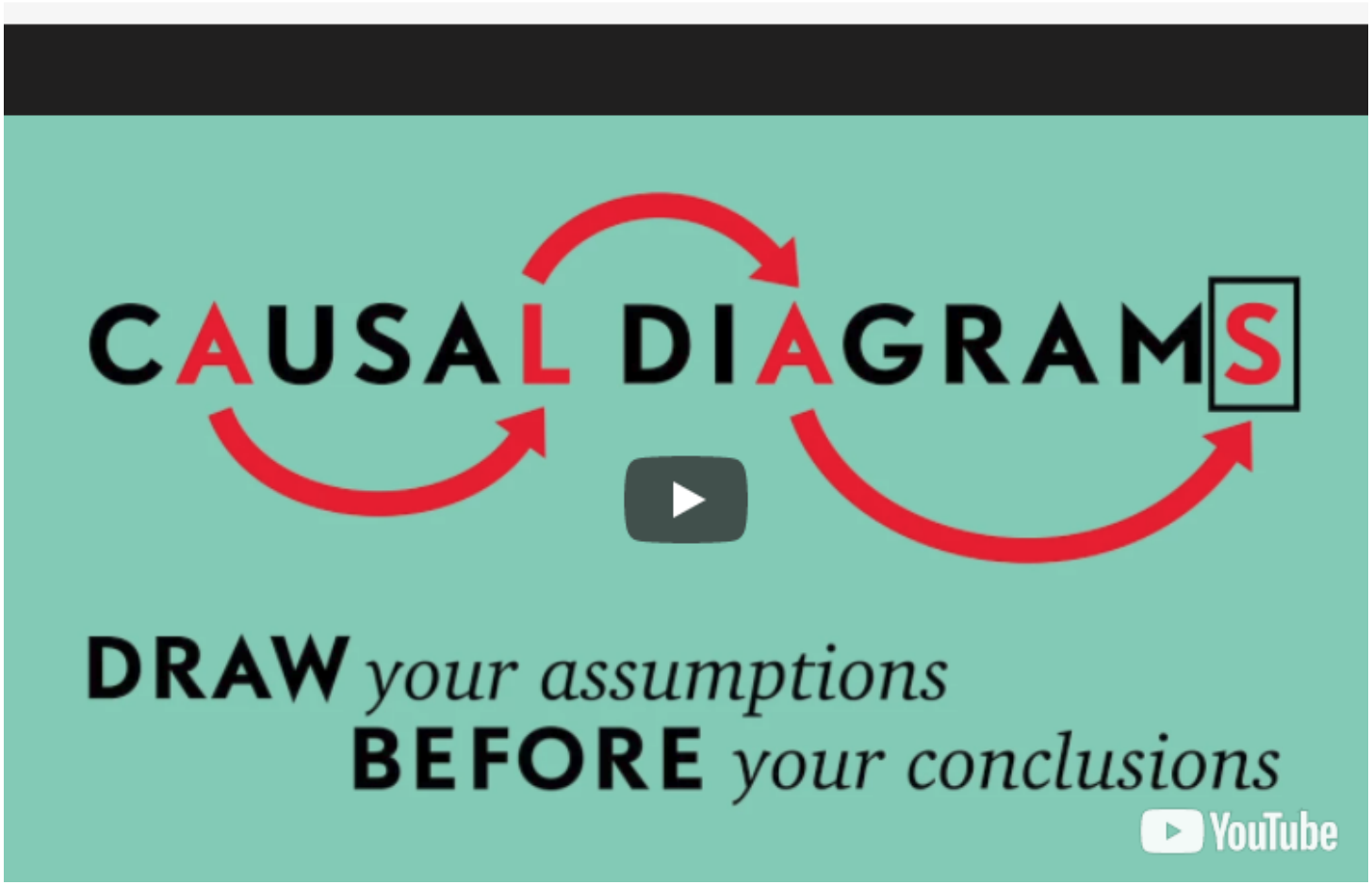
Causal graphs version 3.0 (Open Intro)



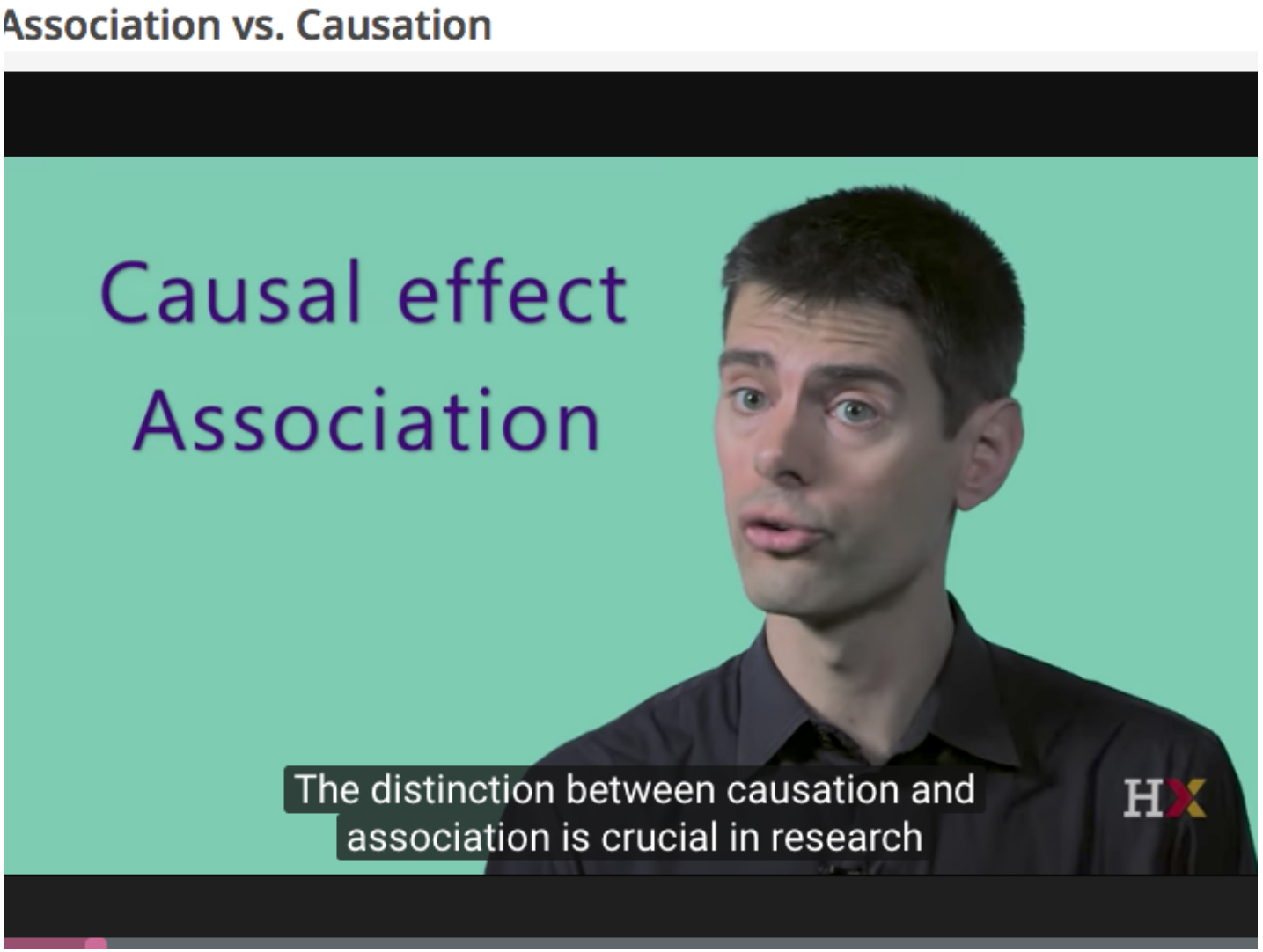
Sun exposure is what is called a **confounding variable**,[13] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

version 3.0: EdX Causal Diagrams (Draw Your Assumptions Before Your Conclusions course)

2. What is a DAG?



version 3.0: EdX Causal Diagrams (Draw Your Assumptions Before Your Conclusions course)



version 3.0: EdX Causal Diagrams (Draw Your Assumptions Before Your Conclusions course)

1. Causal DAGs
2. Confounding
3. Selection Bias
4. Measurement Bias and putting it all together

Learning outcomes (EPI524 Hernan and Swanson)

At the completion of the course, you will be able to:

1. Explain why models are necessary for confounding control
2. Control for confounding using various modeling approaches
3. Identify the relative advantages and disadvantages of each modeling approach
4. Recognize and formulate well defined questions concerning causal effects

EPI524 Description (Hernan and Swanson)

EPI524 describes models for confounding control (or adjustment), their application to epidemiologic data, and the assumptions required to endow the parameter estimates with a causal interpretation. The course introduces students to two broad sets of methods for confounding control: methods that require measuring and appropriately adjusting for confounders, and methods that do not require measuring the confounders.

My proposal (intro stat)

- Include multivariate thinking as part of descriptive statistics
- Include multiple regression as part of that introduction
- Address confounding by stratification and multiple regression control
- Introduce idea of a causal graph
- (Avoid paralysis and paranoia re: "other factors")

My proposal (second course)

- Extend ideas of causal graph and formalize causal inference
- Incorporate Hernan EdX course to save on class time
- Focus on what causal graphs tell us to do in terms of multiple regression, exposure to one or more other methods to address confounding

Making this happen

- Design and confounding are arguably the most important statistical foundation topics (after the concept of variability, now included in K-12 curriculum)
- Rich and sophisticated literature on causal inference now exists
- New curricular models and materials have been created (more needed)
- Need to rethink how we integrate this material into our courses

Teaching multivariate thinking and confounding

1. what are we currently teaching?
2. motivating multivariate examples
3. confounding 101 and 201
4. closing thoughts: next steps

2.6 The Question of Causation*

In many studies of the relationship between two variables, the goal is to es
tablish that changes in the explanatory variable *cause* changes in the respons
variable. Even when a strong association is present, the conclusion that thi
association is due to a causal link between the variables is often hard to find
What ties between two variables (and others lurking in the background) ca

*This section is optional.

Closing thought: need to avoid paralysis

- Rethink our key topics for introductory and intermediate courses
- Ensure that students don't get stuck (conclude they can't make any headway if data don't arise from a randomized trial)
- Teach (modern) design early and often
- Leverage free resources (e.g., Hernan's EdX course) to provide basic background (and free up class time?)
- Reinforce key aspects (observational data vs. randomized trials) when we teach inference
- Teach techniques to move beyond two-sample t-test (stratification and multiple regression)
- Make room by simplifying (what if all datasets were $n > 100$? what if p-values were de-emphasized?)

Multivariate thinking and the introductory statistics and data science course: preparing students to make sense of observational data

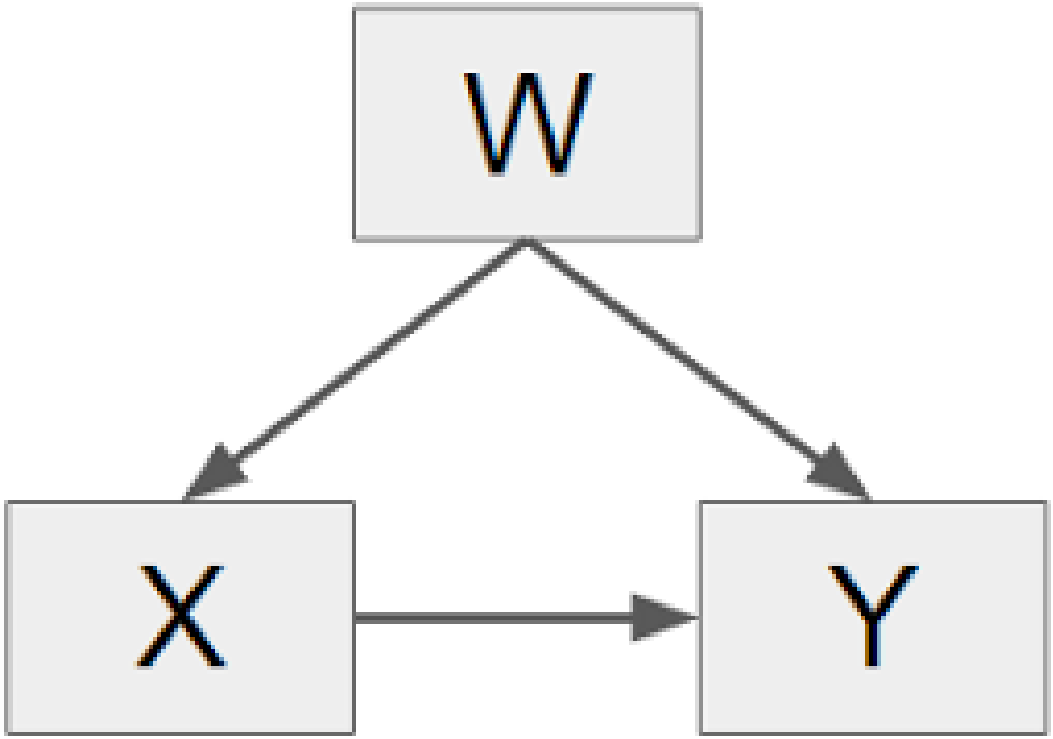
Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

JSM, July 31, 2018

nhorton@amherst.edu
https://www.github.com/Amherst-Statistics/JSM2018

Coffee (X), Cancer (Y), and Smoking (W)



Coffee (X), Cancer (Y), and Smoking (W)