

# 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](mailto:nhorton@amherst.edu)

<https://www.github.com/Amherst-Statistics/JSM2018>

# Thanks and acknowledgments

- Project MOSAIC: Danny Kaplan (Macalester College), Randy Pruim (Calvin College), Ben Baumer (Smith College), and Johanna Hardin (Pomona College)
- Many people working on Causal Inference (Jennifer Hill, Judea Pearl, Richard Scheines, Jamie Robins, Hernan, Rubin, Tyler Vanderweele, Chris Winship, Stephen Morgan, Eric Tchetgen Tchetgen)
- 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 the right things in our introductory statistics and data science courses?

# Motivation

- Are we teaching the right things in our introductory statistics and data science courses?
- OR are students leaving our courses with a form of observational data paralysis?
- Why are we hiding the power of statistics?

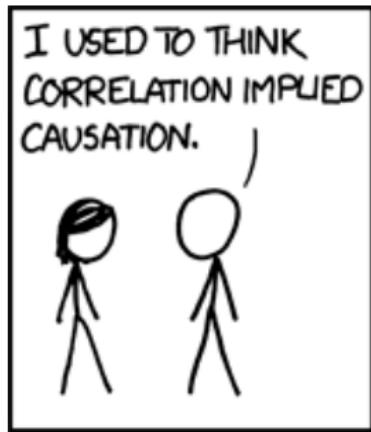


- What evidence do we have that smoking causes lung cancer?



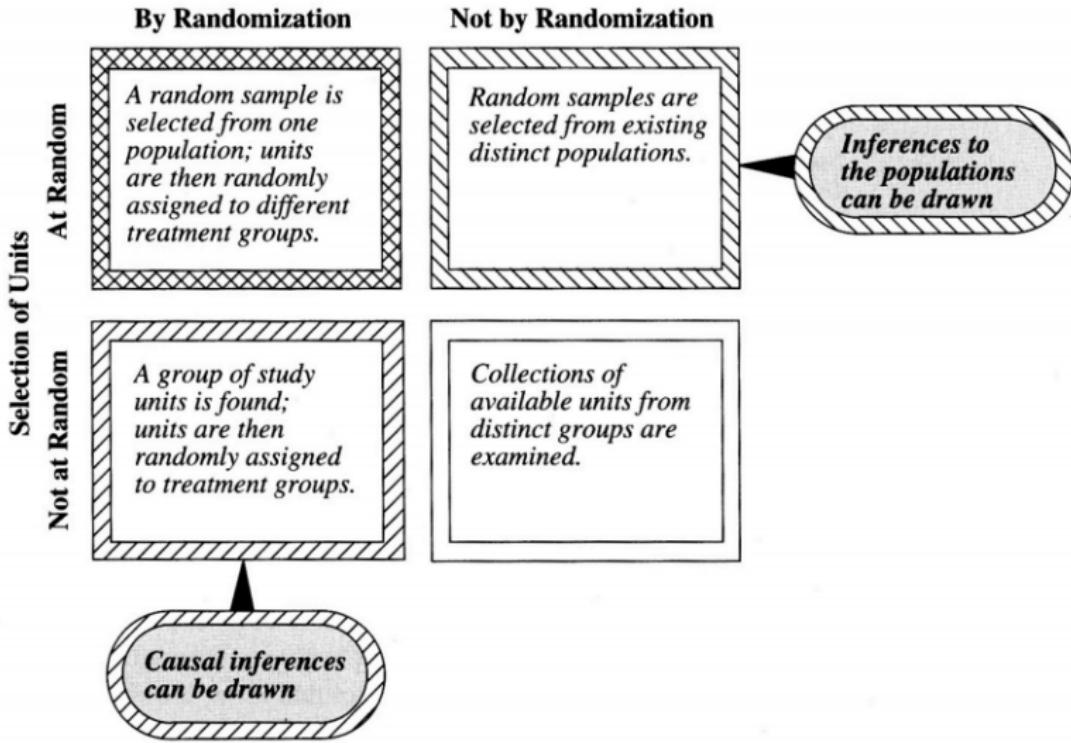
- While clinical trials are wonderful, we live in a world of 'found' data.
- "It is not that I believe an experiment is the only proper setting for discussing causality, but I do feel that it is the simplest such setting" - Holland (1986)

# Do we teach in a way that encourages paralysis?



# Paralysis? (Ramsay and Schafer, Sleuth)

## Allocation of Units to Groups



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?)

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

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

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)

- 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 motivating multivariate examples (Kari Lock Morgan approach via visualization in the prior talk) and multiple regression

# 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 motivating multivariate examples and multiple regression
- ② 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

Jessica M. Utts

Seeing Through  
**Statistics**

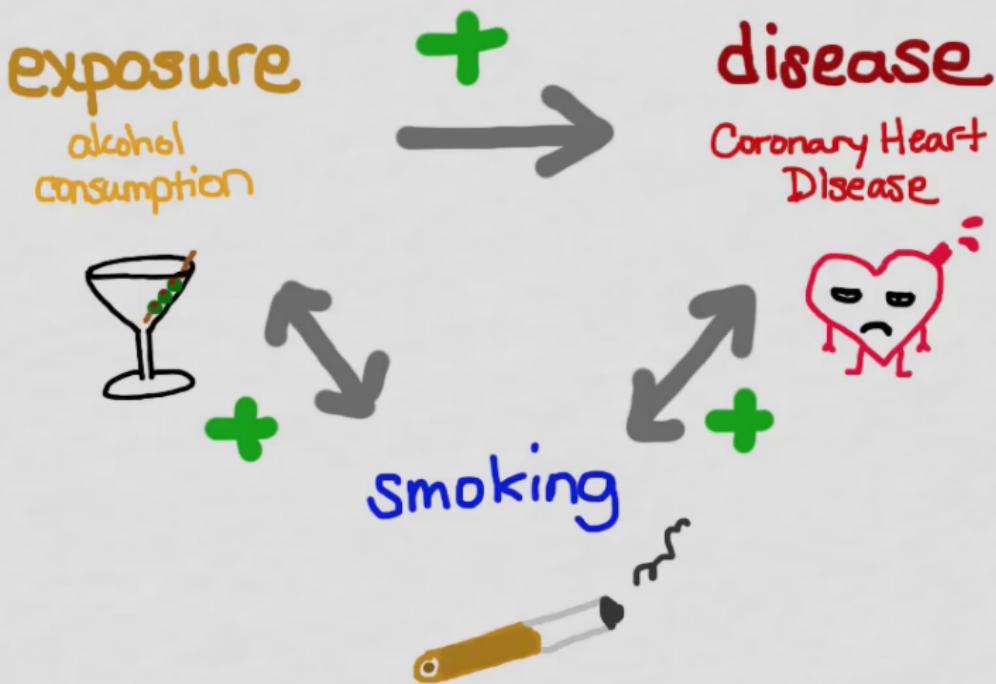
Fourth Edition



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.

## confounding Variable

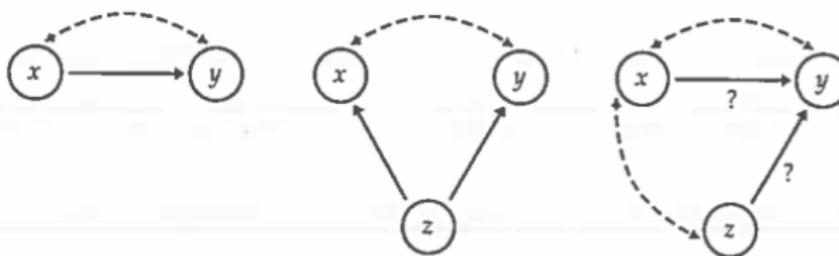


*Confounding is a ubiquitous bias that arises when non-comparable groups are compared. It is one of the greatest threats to valid causal inferences from observational data. Therefore, controlling for confounding is a fundamental component of epidemiologic research.*

## 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)

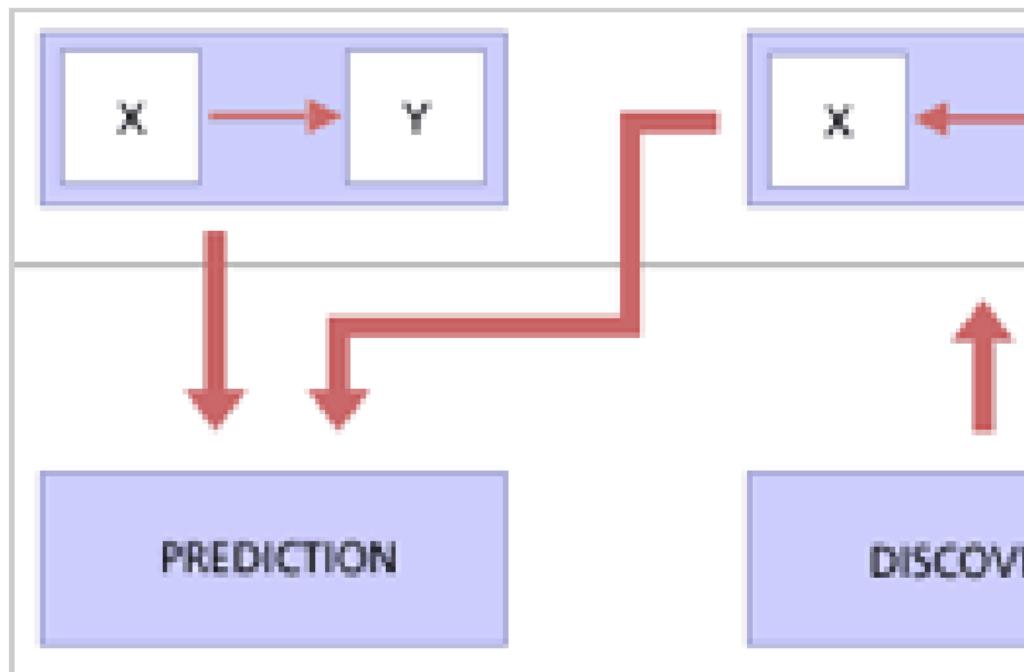
Common response  
(b)

Confounding  
(c)

# What to teach?

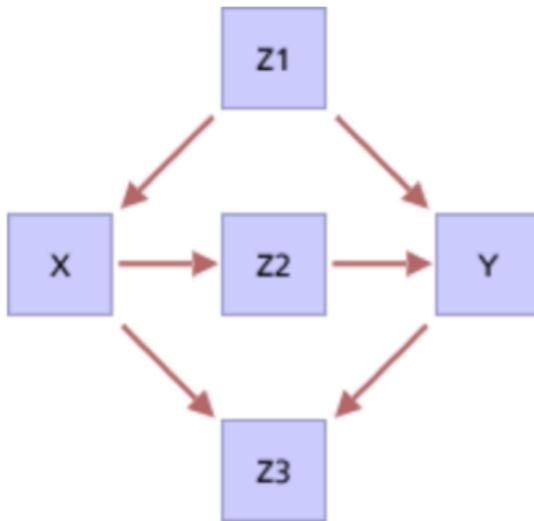
We need to go beyond these informal definitions...

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



- $X$  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 variables 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)

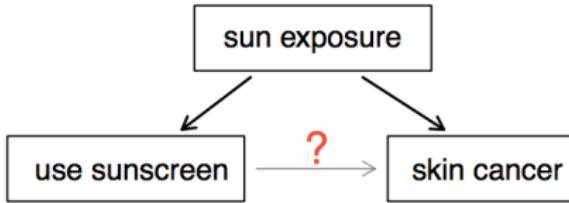
## Example

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 (Open Intro)

20

CHAPTER 1. INTRODUCTION TO DATA



Sun exposure is what is called a **confounding variable**,<sup>13</sup> 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.

## 6.6 The structure of effect modification



Figure 6.11

Identifying potential confounders  
use our causal diagram to  
association between M and Y  
to illustrate the concept of effect modification.

Suppose here we want to  
identify the average causal  
that there is no confounding.  
Computing the average causal  
association is done by calculating  
 $\Pr [Y = 1|A = 1] - \Pr [Y = 1|A = 0]$

## 7.1 The structure of confounding

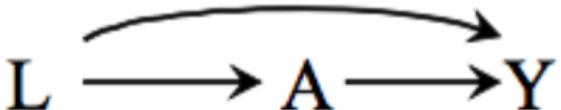


Figure 7.1

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

# EdX Causal Diagrams (Draw Your Assumptions Before Your Conclusions course)

## Association vs. Causation

Causal effect  
Association



The distinction between causation and association is crucial in research



# 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

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

# One proposal (INTRODUCE: intro stat)

- Include multivariate thinking as part of descriptive statistics and visualization
- Include multiple regression (even just as a descriptive statistic)
- Address confounding by stratification and multiple regression control
- INTRODUCE idea of a causal graph
- But no D-separation!
- (Avoid paralysis and paranoia re: “other factors” )

# One proposal (REINFORCE: second course)

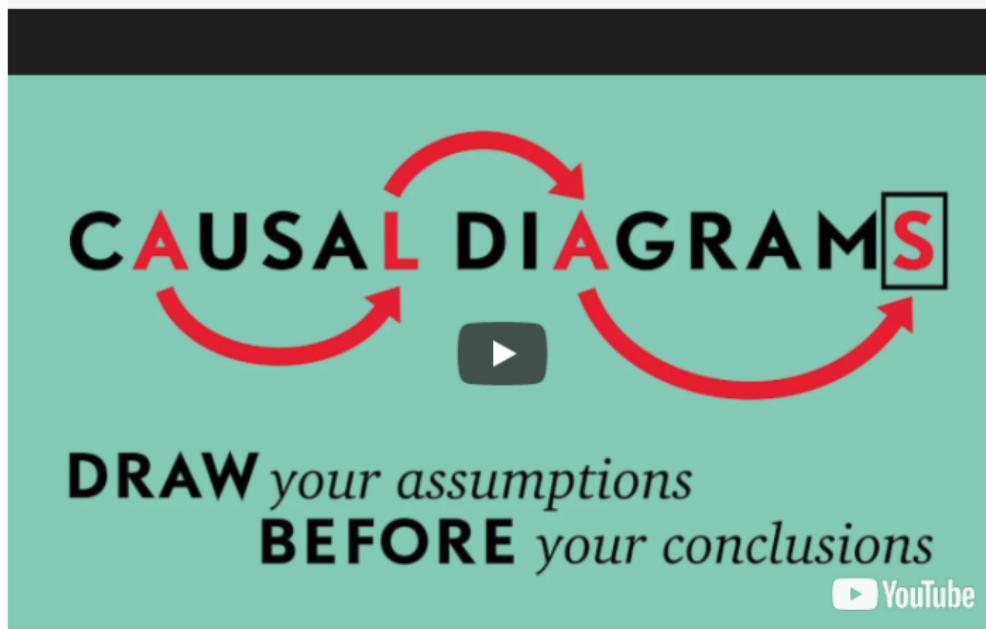
- REINFORCE ideas of causal graph and formalize causal inference
- Incorporate Hernan EdX course to save on class time
- Focus on what causal graphs tell us to do in terms of multiple regression, exposure to one or more other methods to address confounding

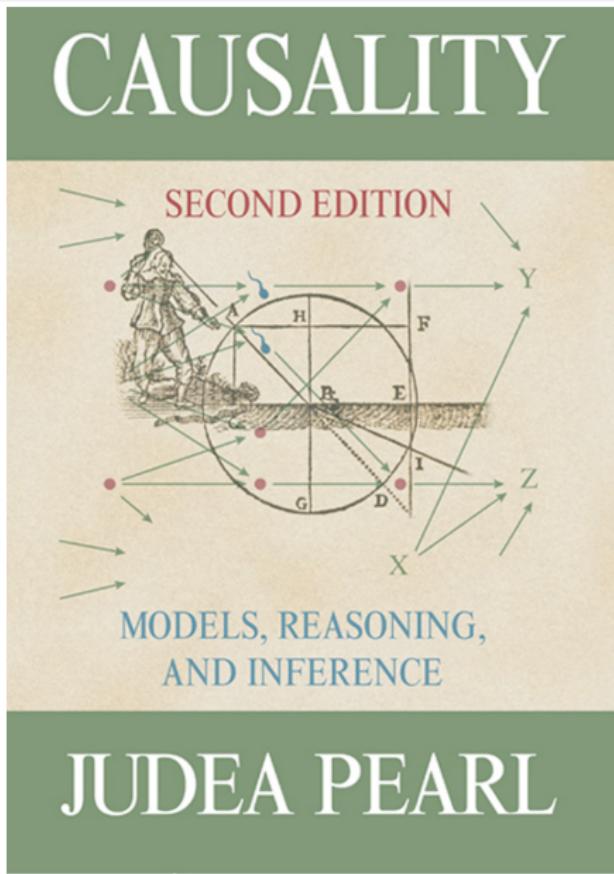
# One proposal (MASTER: Capstone and other courses)

- Work with students to MASTER the idea of confounding and causal inference in later courses and projects

# Need for faculty development

## 2. What is a DAG?



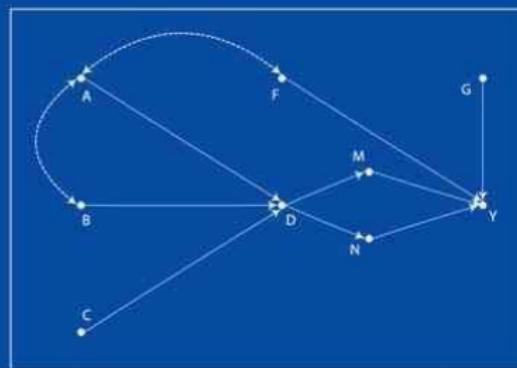


Pearl, second edition (2009)

# Need for faculty development

Winship and Morgan, second edition (2015)

## ANALYTICAL METHODS FOR SOCIAL RESEARCH



## Counterfactuals and Causal Inference

Methods and Principles for Social Research

SECOND EDITION

STEPHEN L. MORGAN  
CHRISTOPHER WINSHIP

- ① motivating multivariate examples and multiple regression
- ② confounding 101 and 201
- ③ closing thoughts: next steps

# Closing thoughts

- Design and confounding are arguably the most important statistical foundation topics
- Builds on precursors now part of the 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

## Closing thoughts: need to avoid paralysis

- Rethink role of causal inference (Introduce, Reinforce, and Master)
- Ensure that students aren't paralyzed by intro!
- Teach techniques that move beyond two-sample t-test (stratification and multiple regression)
- Leverage free resources (e.g., Hernan's EdX course) to provide basic background (and free up class time?)
- Make room by simplifying (what if all datasets were  $n > 100$ ? what if p-values were de-emphasized? Build on K-12 exposure.)

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

Nicholas J. Horton

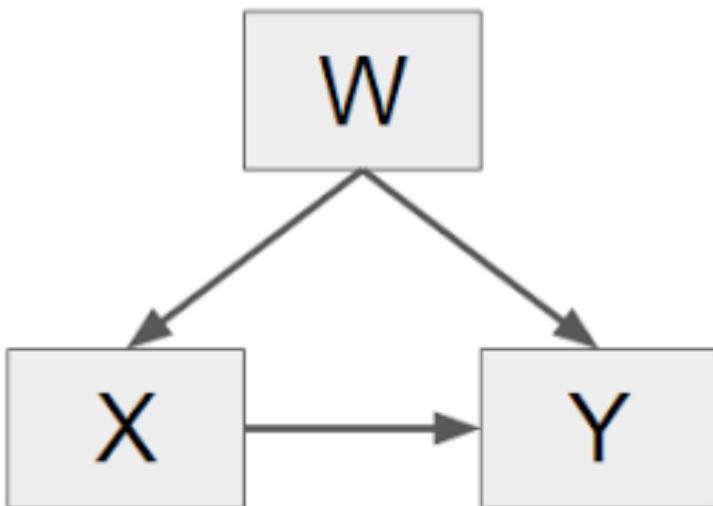
Department of Mathematics and Statistics  
Amherst College, Amherst, MA, USA

JSM, July 31, 2018

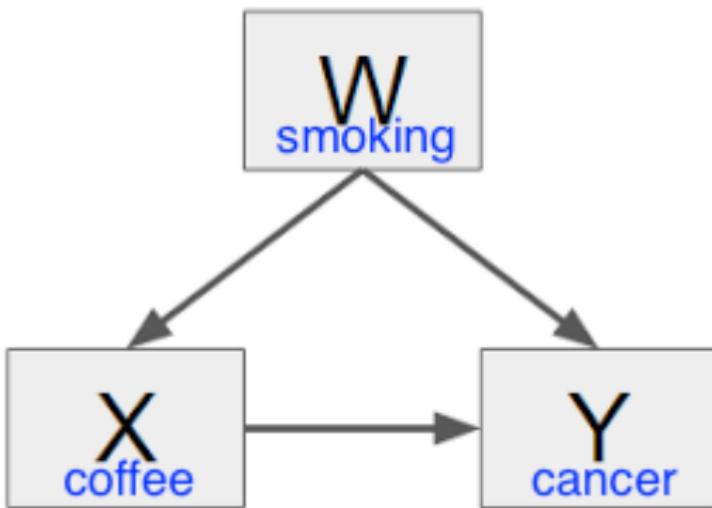
[nhorton@amherst.edu](mailto:nhorton@amherst.edu)

<https://www.github.com/Amherst-Statistics/JSM2018>

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



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



*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.*