

# 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

Babson College, April 30, 2018

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

<http://nhorton.people.amherst.edu>

# 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

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?

OR are we hiding the power of statistics behind our bushel basket?



- 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 multivariate thinking?

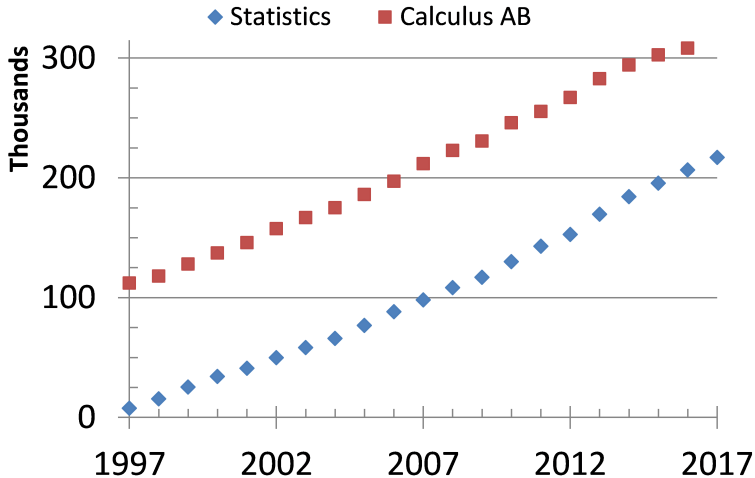
*As we begin the 21st century, the introductory statistics course appears healthy, with its emphasis on real examples, data production, and graphics for exploration and assumption-checking. Without doubt this emphasis marks a major improvement over introductory courses of the 1960s, an improvement made possible by the vaunted “computer revolution.” Nevertheless, I argue that despite broad acceptance and rapid growth in enrollments, the consensus curriculum is still an **unwitting prisoner of history** (Cobb, TISE 2007).*



# Teaching multivariate thinking and confounding

- 1 what are we currently teaching?

## AP Exams



Data Source: College Board

Four broad themes:

- 1 Exploring Data: Describing patterns and departures from patterns
- 2 Sampling and Experimentation: Planning and conducting a study
- 3 Anticipating Patterns: Exploring random phenomena using probability and simulation
- 4 Statistical Inference: Estimating population parameters and testing hypotheses

# AP Stats: Exploring data (20-30%)

- 1 Univariate graphics
- 2 Univariate summaries
- 3 Comparing univariate distributions
- 4 Exploring bivariate data
- 5 Exploring categorical data

# AP Stats: Sampling and experimentation (10-15%)

- 1 Methods of data collection
- 2 Planning and conducting surveys
- 3 Planning and conducting experiments
- 4 Generalizability of results and types of conclusions

# AP Stats: Anticipating patterns (20-30%)

- 1 Probability
- 2 Rules for random variables
- 3 Normal tables(!)
- 4 Sampling distributions
- 5 Other distributions (t,  $\chi^2$ )

# AP Stats: Statistical inference (30-40%)

- 1 Estimation (point estimators and confidence intervals)
- 2 Tests of significance

*Currently, the graphing calculator is the only computational aid that is available to students for use as a tool for data analysis on the AP Exam (AP Stats course description).*

Relatively modern curriculum from circa 1996, but little beyond bivariate comparisons



MA 113 (Boston University) Basic concepts of estimation and tests of hypotheses, ideas from probability; one-, two-, and multiple-sample problems. Applications in social sciences. Primarily for students in the social sciences who require a one-semester introduction to statistics.

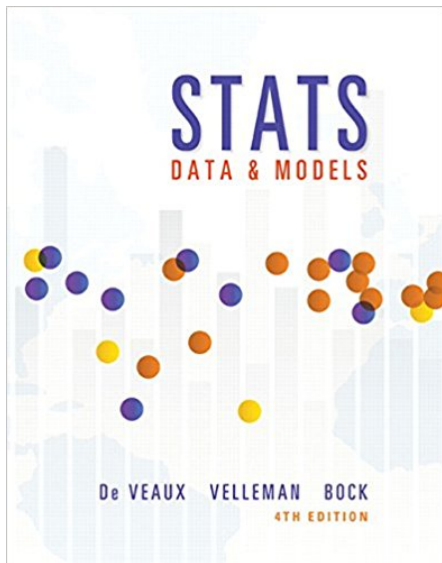
MATH125 (UMass/Boston) Topics include introductory statistics, covering descriptive statistics; introductory probability sufficient to enable development of inferential statistics; and inferential statistics.

**STAT101 (Wellesley)** An introduction to the fundamental ideas and methods of statistics for analyzing data. Topics include descriptive statistics, basic probability, inference, and hypothesis testing. Emphasis on understanding the use and misuse of statistics in a variety of fields, including medicine and both the physical and social sciences.

STAT100 (Harvard Stat) Introduction to key ideas underlying statistical and quantitative reasoning. Topics covered: methods for organizing, summarizing and displaying data; elements of sample surveys, experimental design and observational studies; methods of parameter estimation and hypothesis testing in **one- and two-sample problems**; regression with **one or more predictors**; correlation; and analysis of variance. Explores applications in a wide range of fields, including the social and political sciences, medical research, and business and economics.

**BST201 (Harvard Chan)** Covers basic statistical techniques that are important for analyzing data arising from epidemiology, environmental health and biomedical and other public health-related research. Major topics include descriptive statistics, elements of probability, introduction to estimation and hypothesis testing, nonparametric methods, techniques for categorical data, **regression analysis**, analysis of variance, and elements of study design. Applications are stressed. Designed for students desiring more emphasis on theoretical developments.

QTM1000 (Babson College) The course introduces the necessary core quantitative methods ... Statistical software and the use of spreadsheets are integrated throughout so that students better comprehend the importance of using modern technological tools for effective model building and decision-making. About two thirds of the course is data-oriented, exposing students to basic statistical methods, their conceptual underpinning, such as variability and uncertainty, and their use in the real world. Topics include data collection, descriptive statistics, elementary probability rules and distributions, sampling distributions, and basic inference. (Followed by QTM1010, linear regression [multivariate?] and QTM2000, more modeling.)



- Chapter 2: “When averages are taken across different groups, they can appear to contradict the overall averages. This is known as *Simpson’s paradox*”
- Chapter 12: introduce confounding (in the context of clinical trials)
- Chapter 28 (page 817) multiple regression
- (material slated to appear earlier in future editions...)



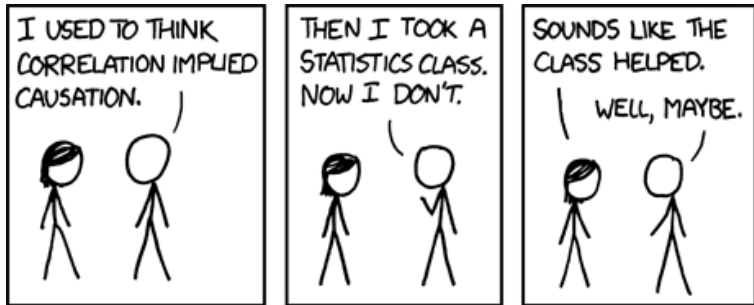
SECOND EDITION  
 PRINCIPLES  
 OF  
 BIostatISTICS

*The CONQUEST OF PESTILENCE in NEW YORK CITY*  
 —As Shown by the Death Rate as Recorded in the Official Boards of the Department of Health.

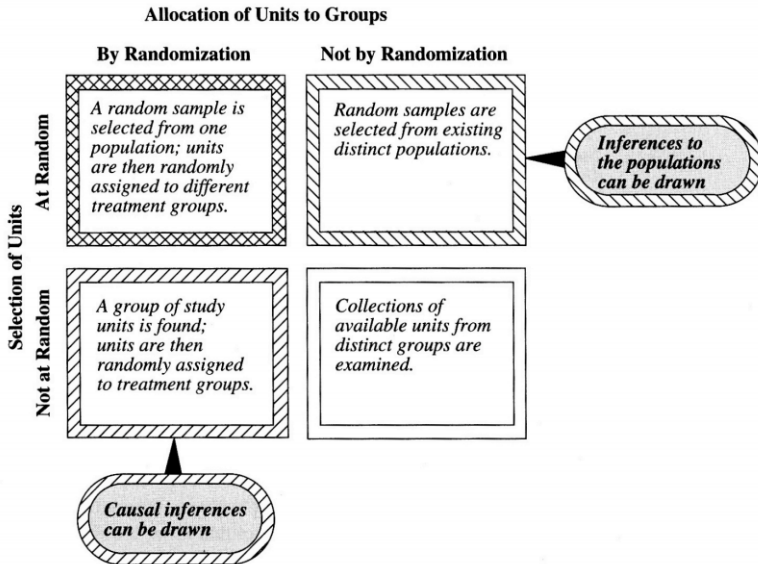


- Chapter 4: Confounding defined (page 71, standardization as an approach to address p. 84–89)
- Chapter 16: Simpson's paradox introduced via multiple 2x2 tables (page 374 of 525)
- Chapter 19: multiple regression to move beyond bivariate questions
- Why wait?

# 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.

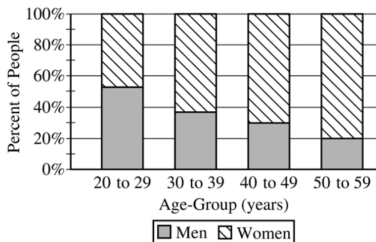
(But what about possible confounders?)



## 2017 AP® STATISTICS FREE-RESPONSE QUESTIONS

5. The table and the bar chart below summarize the age at diagnosis, in years, for a random sample of 207 men and women currently being treated for schizophrenia.

	Age-Group (years)				
	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46	40	21	12	119
Men	53	23	9	3	88
Total	99	63	30	15	207



Do the data provide convincing statistical evidence of an association between age-group and gender in the diagnosis of schizophrenia?

# OpenIntro Statistics

Third Edition



David M Diez  
Christopher D Barr  
Mine Çetinkaya-Rundel

---

# 3

---

**6.34 Prenatal vitamins and Autism.** Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptual period).<sup>57</sup>

		<i>Autism</i>		Total
		Autism	Typical development	
<i>Periconceptual prenatal vitamin</i>	No vitamin	111	70	181
	Vitamin	143	159	302
	Total	254	229	483

- State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.
- Complete the hypothesis test and state an appropriate conclusion. (Reminder: Verify any necessary conditions for the test.)
- A New York Times article reporting on this study was titled “Prenatal Vitamins May Ward Off Autism”. Do you find the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.<sup>58</sup>

- 1 Since the p-value  $< \alpha$ , we reject  $H_0$ . There is strong evidence of a difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.

- 1 Since the p-value  $< \alpha$ , we reject  $H_0$ . There is strong evidence of a difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.
- 2 The p-value is small and we reject  $H_0$ . The data provide convincing evidence to suggest that autism and vitamin use in women are associated.

- 1 Since the p-value  $< \alpha$ , we reject  $H_0$ . There is strong evidence of a difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.
- 2 The p-value is small and we reject  $H_0$ . The data provide convincing evidence to suggest that autism and vitamin use in women are associated.
- 3 Yes, this is an observational study. Based on this study we can't deduce that taking vitamins leads to less autism. There may be other factors, lurking variables, that may cause autism and be associated with vitamin use.

# (Non-scientific) survey of isolated statisticians and Stat Ed section members

question: “what assumptions do you have students check when using the two sample t-test?”

# (Non-scientific) survey of isolated statisticians and Stat Ed section members

question: “what assumptions do you have students check when using the two sample t-test?”

representative answer: (instructor using Gould and Ryan [first edition])

- 1 Randomness in the data collection process (either **random samples** or **experiment**)
- 2 Independent samples
- 3 Either normal looking samples or sample sizes larger than 25



# (Non-scientific) survey of isolated statisticians and Stat Ed section members

**question:** “what assumptions do you have students check when using the two sample t-test?”

**representative answer:** (instructor using Gould and Ryan [first edition])

- 1 Randomness in the data collection process (either **random samples** or **experiment**)
- 2 Independent samples
- 3 Either normal looking samples or sample sizes larger than 25

What about possible confounders? (Only one other respondent out of more than 20 mentioned “random assignment”: almost all emphasis was on technical conditions).

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

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

We conclude that there is statistically significant evidence of an association between CAT level and CHD risk in these data.

(assuming no confounding, no selection bias, no information bias)

# Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

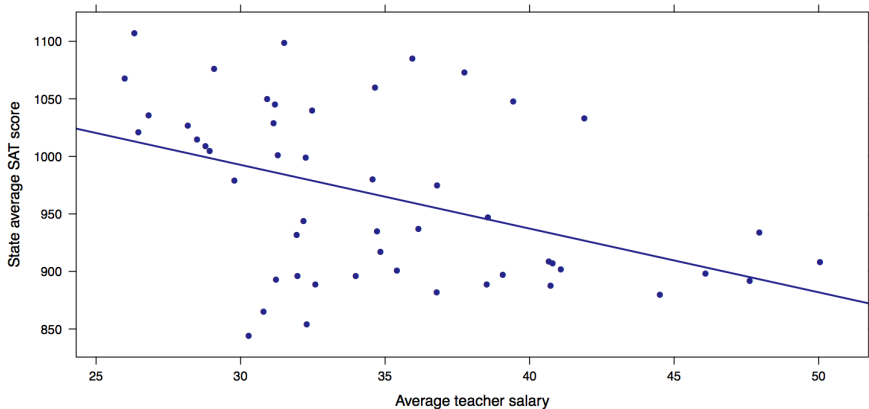
- 1 Teach statistical thinking.
- 2 Focus on conceptual understanding.
- 3 Integrate real data with a context and purpose.
- 4 Foster active learning.
- 5 Use technology to explore concepts and analyze data.
- 6 Use assessments to improve and evaluate student learning.

- 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.

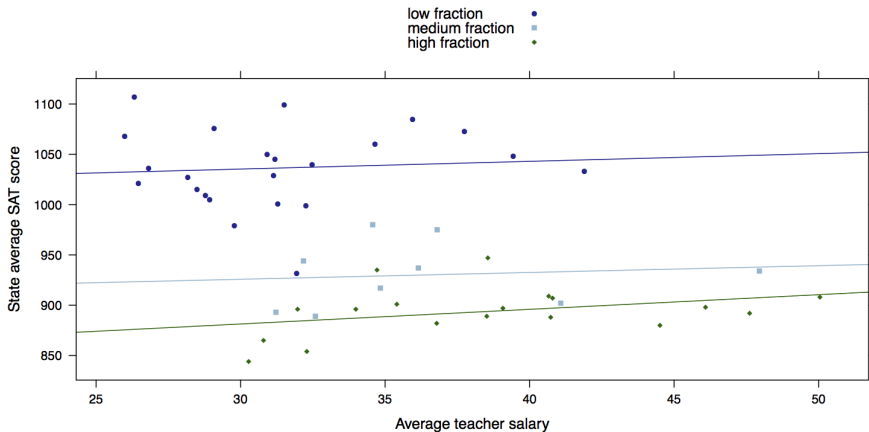
# Teaching multivariate thinking and confounding

- 1 what are we currently teaching?
- 2 motivating multivariate examples

# SAT scores and teacher salaries (state data from 2010)



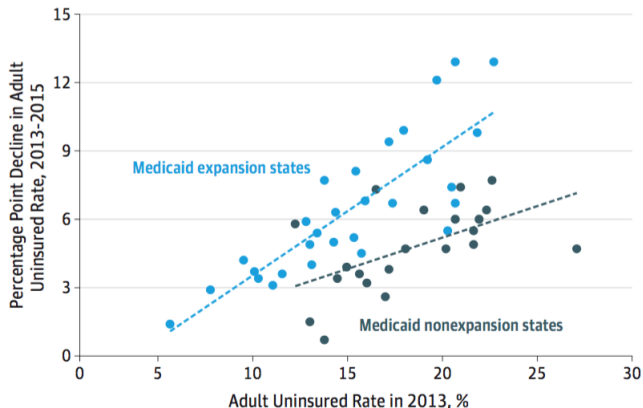
# SAT scores and teacher salaries (state data from 2010)





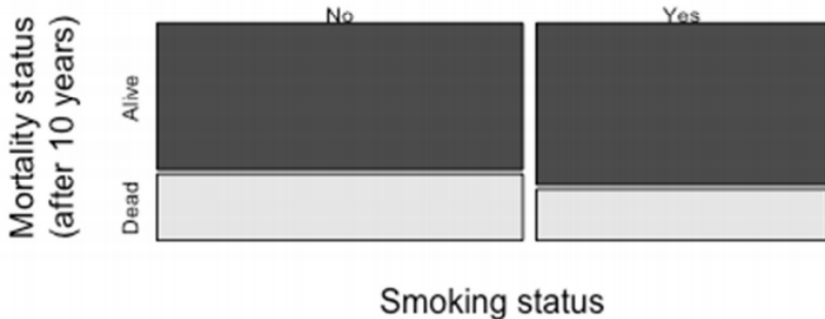
# stratification and/or multiple regression: Obama's 2016 single author JAMA paper

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

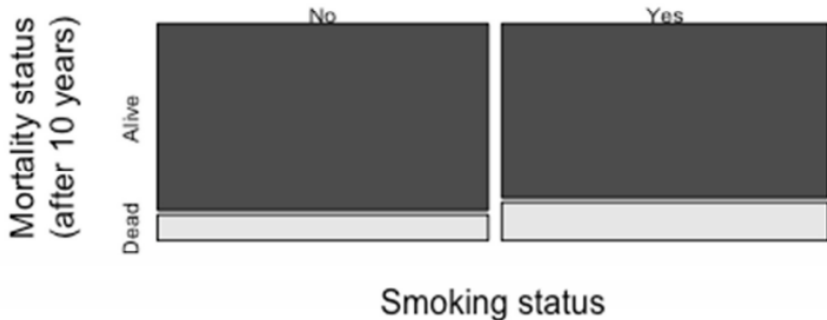


See also "Statistical methods in the NEJM" (2007)

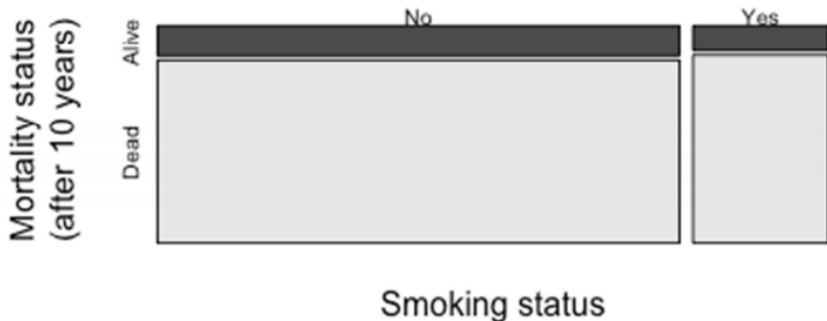
## Association of smoking and mortality



## Results for 18-64 year olds at baseline



## Results for those 65+ years old at baseline



# Kidney stones (Wikipedia Simpson's Paradox)

	<b>Treatment A</b>	<b>Treatment B</b>
<b>Small stones</b>	<i>Group 1</i> <b>93% (81/87)</b>	<i>Group 2</i> <b>87% (234/270)</b>
<b>Large stones</b>	<i>Group 3</i> <b>73% (192/263)</b>	<i>Group 4</i> <b>69% (55/80)</b>
<b>Both</b>	<b>78% (273/350)</b>	<b>83% (289/350)</b>

# How to handle more than two variables?

- “Data Viz on Day One” (TISE, <http://escholarship.org/uc/item/84v3774z>)

# How to handle more than two variables?

- “Data Viz on Day One” (TISE, <http://escholarship.org/uc/item/84v3774z>)
- stratification

# 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)



# 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)

# 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)

Jessica M. Utts

## Seeing Through Statistics

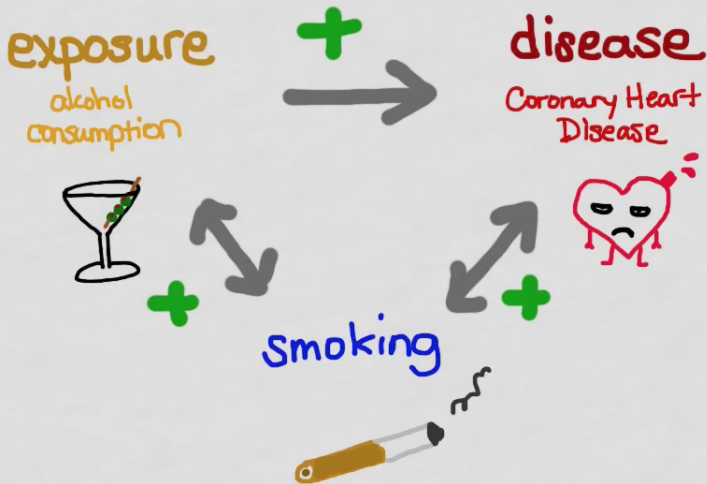
Fourth Edition



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.*



*In statistics, a confounder (also confounding variable or confounding factor) is a variable that influences both the dependent variable and independent variable causing a spurious association. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations.*



**WIKIPEDIA**  
The Free Encyclopedia

[Main page](#)

[Contents](#)

[Featured content](#)

[Current events](#)

[Random article](#)

[Donate to Wikipedia](#)

[Wikipedia store](#)

[Interaction](#)

[Help](#)

[About Wikipedia](#)

[Community portal](#)

[Recent changes](#)

[Contact page](#)

[Bikestats](#)



[Talk](#)

[Sandbox](#)

[Preferences](#)

[Beta](#)

[Watchlist](#)

[Contributi](#)

[Article](#)

[Talk](#)

[Read](#)

[Edit source](#)

[New section](#)

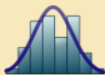
[View history](#)

[More](#) ▾

## Talk:Confounding

From Wikipedia, the free encyclopedia

[157.27.198.235 \(talk\)](#) 11:37, 16 September 2016 (UTC)



This article is within the scope of the **WikiProject Statistics**, a



[Statistics portal](#)

collaborative effort to improve the coverage of [statistics](#) on Wikipedia. If you would like to participate, please visit the [project page](#) or join the [discussion](#).

**C**

This article has been rated as **C-Class** on the [quality scale](#).

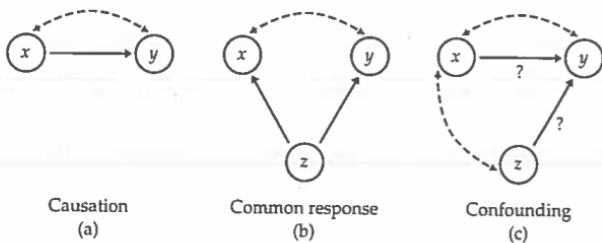
**High**

This article has been rated as **High-importance** on the [importance scale](#).

## 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  $y$ .

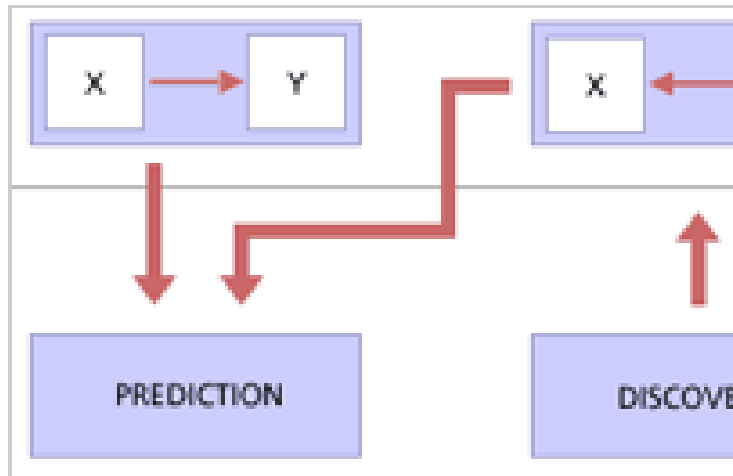
Items 1 and 2 in Example 2.42 are examples of direct causation. *Even when direct causation is present, very often it is not a complete explanation of an association between two variables.* The best evidence for causation comes from experiments that actually change  $x$  while holding all other factors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



# What to teach?

We need to go beyond these informal definitions...

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



Causal

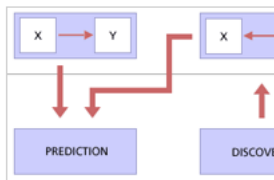
Over

This

reas

mak

soc



## Causal & Statistical Reasoning [\[Enter Course\]](#)

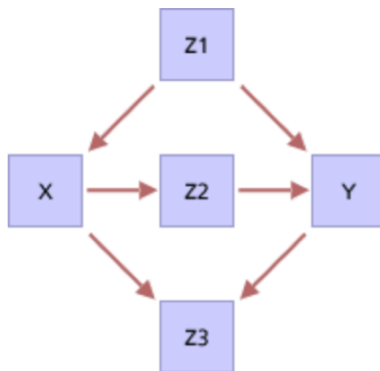
### Overview:

This course provides an introduction to causal and statistical reasoning. After taking this course, students will be better prepared to make rational decisions about their own lives and about matters of social policy. They will be able to assess critically—even if informally—claims that they encounter during discussions or when considering a news article or report. A variety of materials are presented, including Case Studies where students are given the opportunity to examine a causal claim, and the Causality Lab, a virtual environment to simulate the science of causal discovery. Students have frequent opportunities to check their understanding and practice their skills.

- 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)

## Definition [\[ edit | edit source \]](#)

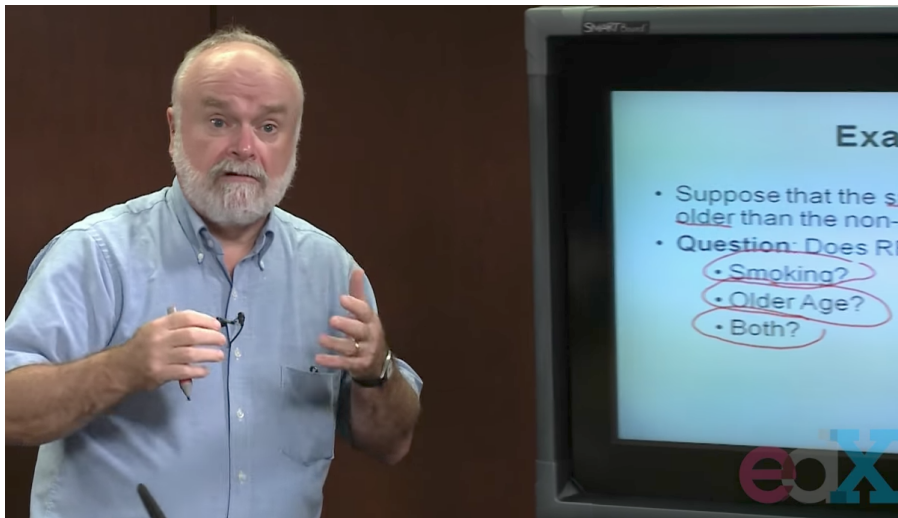
Confounding is defined in terms of the data generating model (as in the Figure above). Let  $X$  be some independent variable,  $Y$  some dependent variable. To estimate the effect of  $X$  on  $Y$ , the statistician must suppress the effects of extraneous variables that influence both  $X$  and  $Y$ . We say that,  $X$  and  $Y$  are confounded by some other variable  $Z$  whenever  $Z$  is a **cause** of both  $X$  and  $Y$ .

Let  $P(\mathbf{y} \mid \text{do}(\mathbf{x}))$  be the probability of event  $Y = y$  under the hypothetical intervention  $X = x$ .  $X$  and  $Y$  are not confounded if and only if the following holds:

$$P(\mathbf{y} \mid \text{do}(\mathbf{x})) = P(\mathbf{y} \mid \mathbf{x}) \tag{1}$$

for all values  $X = x$  and  $Y = y$ , where  $P(\mathbf{y} \mid \mathbf{x})$  is the conditional probability upon seeing  $X = x$ . Intuitively, this equality states that  $X$  and  $Y$  are not confounded whenever the observationally witnessed association between them is the same as the association that would be measured in a controlled experiment, with  $x$  randomized.

# Causal graphs version 2.0 (Fran Cook videos)



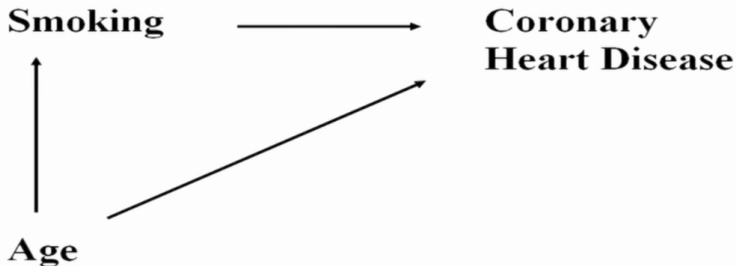
The image shows a man with a grey beard and glasses, wearing a light blue button-down shirt, gesturing with his hands as if speaking. To his right is a large screen displaying a slide. The slide has a light blue background and contains the following text:

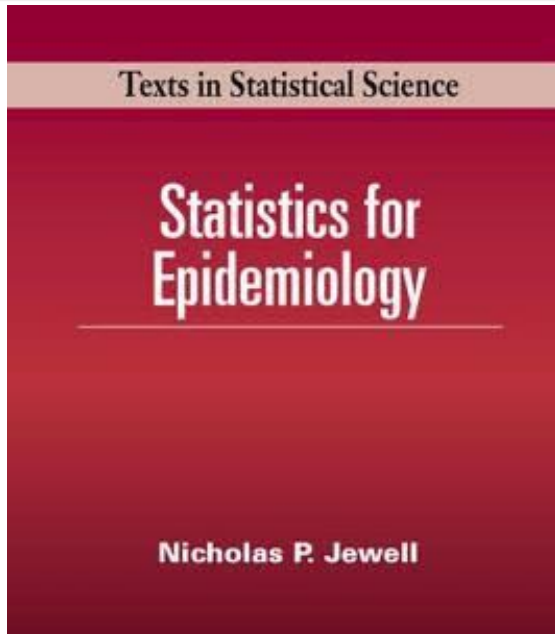
Exa

- Suppose that the s  
older than the non-
- Question: Does R  
• Smoking?  
• Older Age?  
• Both?

The slide also features the edX logo in the bottom right corner.

## Direct Acyclic Graph (DAG)





# CAUSALITY

SECOND EDITION



MODELS, REASONING,  
AND INFERENCE

# JUDEA PEARL

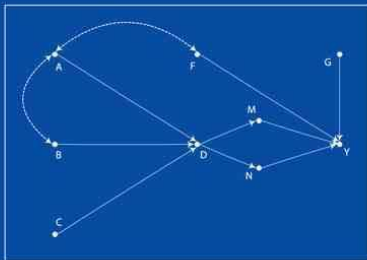
Springer Series in Statistics

Mark J. van der Laan  
Sherri Rose

# Targeted Learning

Causal Inference for Observational  
and Experimental Data

ANALYTICAL METHODS FOR SOCIAL RESEARCH



## Counterfactuals and Causal Inference

Methods and Principles for Social Research

SECOND EDITION

STEPHEN L. MORGAN  
CHRISTOPHER WINSHIP



# EXPLANATION IN CAUSAL INFERENCE

Methods for Mediation and Interaction

TYLER J. VANDERWEELE

## 6.6 The structure of effect modification

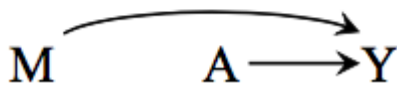


Figure 6.11

Identifying potential  
 use our causal  
 association bet  
 to illustrate th

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

## 7.1 The structure of confounding

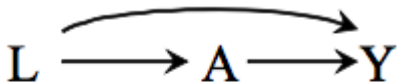
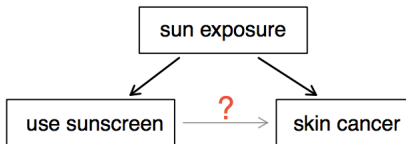


Figure 7.1

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

# Example

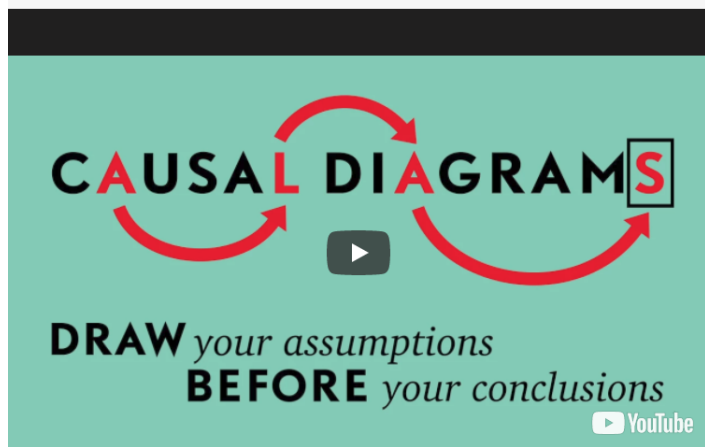
Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?



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.

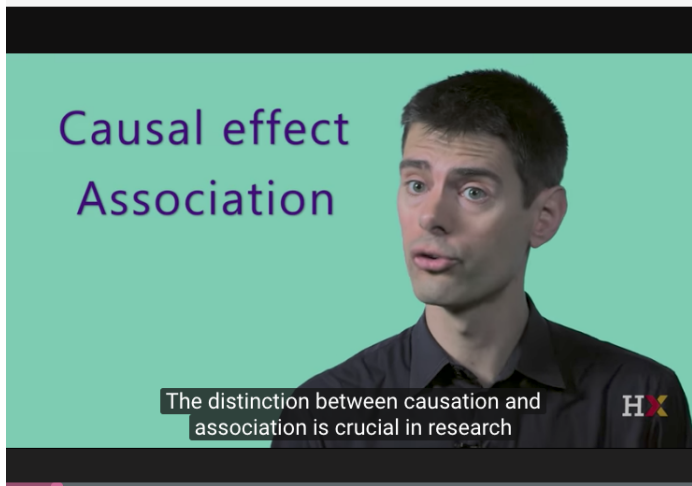
# version 3.0: EdX Causal Diagrams (Draw Your Assumptions Before Your Conclusions course)

## 2. What is a DAG?



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

## Association vs. Causation



Causal effect  
Association

The distinction between causation and association is crucial in research

HX

A video player interface showing a man in a dark shirt speaking. The background is a light teal color. The text 'Causal effect' and 'Association' is displayed in purple. A subtitle at the bottom reads 'The distinction between causation and association is crucial in research'. The HX logo is visible in the bottom right corner of the video frame.

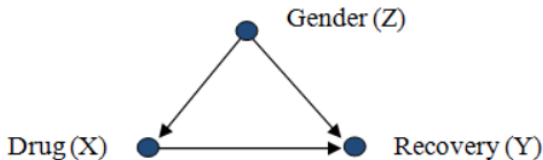
# version 3.0: EdX Causal Diagrams (Draw Your Assumptions Before Your Conclusions course)

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



## Control [\[ edit | edit source \]](#)

Consider a researcher attempting to assess the effectiveness of drug X, from population data in which drug usage was a patient's choice. Data show that gender(Z) differences influence a patient's choice of drug as well as their chances of recovery (Y). In this scenario, gender Z confounds the relation between X and Y since Z is a cause of both X and Y:

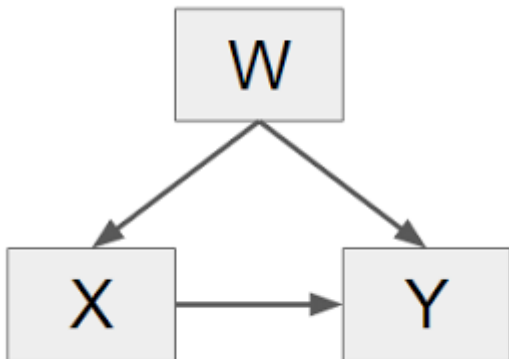


We have that

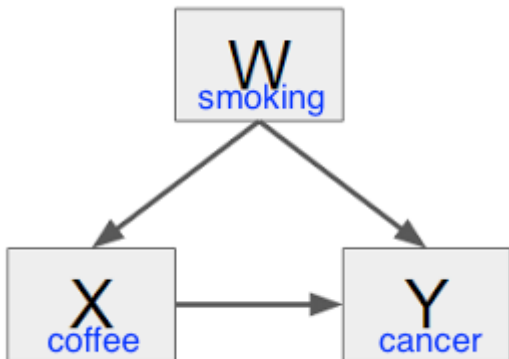
$$P(y \mid do(x)) \neq P(y \mid x) \tag{2}$$

because the observational quantity contains information about the correlation between X and Z, and the interventional quantity does not (since X is not correlated with Z in a randomized experiment). Clearly the statistician desires the unbiased estimate

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



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



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 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.*

*Specifically, the course introduces outcome regression, propensity score methods, the parametric g-formula, inverse probability weighting of marginal structural models, and instrumental variable methods as means for confounding control. The models described in EPI524 are for time-fixed dichotomous exposures and dichotomous, continuous, and failure time (e.g., survival) outcomes.*

# EPI524 Description (Hernan and Swanson)

- Pretty tough sledding
- Can we determine learning outcomes that are more accessible to a broader audience?
- Can we leverage free resources (e.g., Hernan's EdX course) to provide basic background (and free up class time?)

# 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 (Stat 2)

- 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

# My proposal (Capstone and other courses)

- Reinforce and extend these ideas in later courses and projects

Home » Member News, People News

## 2017 Causality in Statistics Award Announced

1 AUGUST 2017

219 VIEWS

NO COMMENT

The American Statistical Association will award the fifth [Causality in Statistics](#) award to [Ilya Shpitser](#), John C. Malone Assistant Professor of Computer Science at The Johns Hopkins University, at the [2017 Joint Statistical Meetings in Baltimore](#).

# Making this happen

- Design and confounding are arguably the most important statistical foundation topics (after the concept of variability)
- 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 (and promulgate the approach)

# 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 establish that changes in the explanatory variable *cause* changes in the response variable. Even when a strong association is present, the conclusion that this 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) can

---

\*This section is optional.

# Next steps

- focus on writing (e.g., KB talk to follow)
- focus on projects (e.g., Katherine, Jessen, and Patricia talks)
- focus on visualization (e.g., Vetricia's talk)
- rethink our key topics for introductory and intermediate courses

# Closing thought: need to avoid paralysis

- 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
- 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

Babson College, April 30, 2018

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

<http://nhorton.people.amherst.edu>

## Big Idea 3: Data and Information

**Data and information facilitate the creation of knowledge.** Computing enables and empowers new methods of information processing, driving monumental change across many disciplines — from art to business to science. Managing and interpreting an overwhelming amount of raw data is part of the foundation of our information society and economy. People use computers and computation to translate, process, and visualize raw data and to create information.

Computation and computer science facilitate and enable new understanding of data and information that contributes knowledge to the world. Students in this course work with data using a variety of computational tools and techniques to better understand the many ways in which data is transformed into information and knowledge.

# AP Computer Science Principles: taught for first time last year

---

## **Enduring Understandings**

(Students will understand that ... )

## **Learning Objectives**

(Students will be able to ... )

---

**EU 3.1** People use computer programs to process information to gain insight and knowledge.

**LO 3.1.1** Find patterns and test hypotheses about digitally processed information to gain insight and knowledge. [P4]

**LO 3.1.3** Explain the insight and knowledge gained from digitally processed data by using appropriate visualizations, notations, and precise language. [P5]

**EK 3.1.3A** Visualization tools and software can communicate information about data.

---

**EK 3.1.3B** Tables, diagrams, and textual displays can be used in communicating insight and knowledge gained from data.

---

**EK 3.1.3C** Summaries of data analyzed computationally can be effective in communicating insight and knowledge gained from digitally represented information.

---

**EK 3.1.3D** Transforming information can be effective in communicating knowledge gained from data.

---

**EK 3.1.3E** Interactivity with data is an aspect of communicating.

# AP Computer Science Principles: 200 page course description

**EU 3.2** Computing facilitates exploration and the discovery of connections in information.

**LO 3.2.1** Extract information from data to discover and explain connections or trends. [P1]

8 of 9      Biologists often attach tracking collars to wild animals. For each animal, the following geolocation data is collected at frequent intervals.

- The time
- The date
- The location of the animal

Which of the following questions about a particular animal could NOT be answered using only the data collected from the tracking collars?

Approximately how many miles did the animal travel in one week?

---

Does the animal travel in groups with other tracked animals?

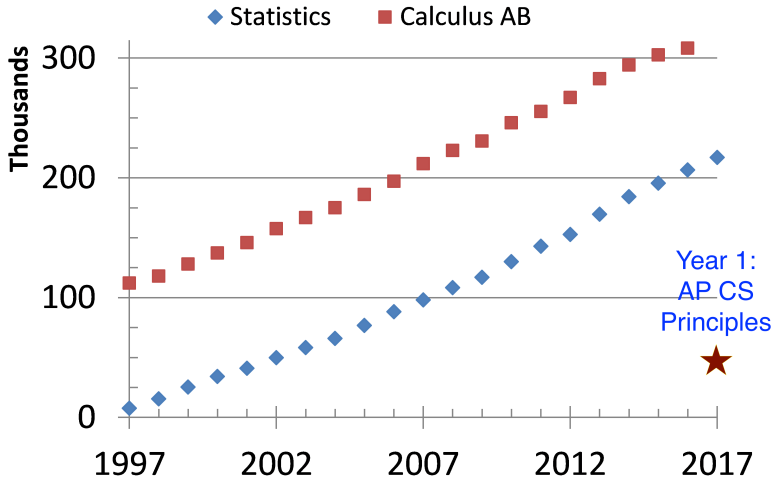
---

Do the movement patterns of the animal vary according to the weather?

---

In what geographic locations does the animal typically travel?

## AP Exams



Data Source: College Board