Avoiding paralysis via multivariate thinking

Nicholas J. Horton

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

USCOTS, May 18, 2017

nhorton@amherst.edu
http://nhorton.people.amherst.edu

Potential paralysis?

- Wild: glimpses of a multivariate world

Potential paralysis?

- Wild: glimpses of a multivariate world
- Question: do we reinforce key aspects of *design* (observational data vs. randomized trials) when we teach inference?

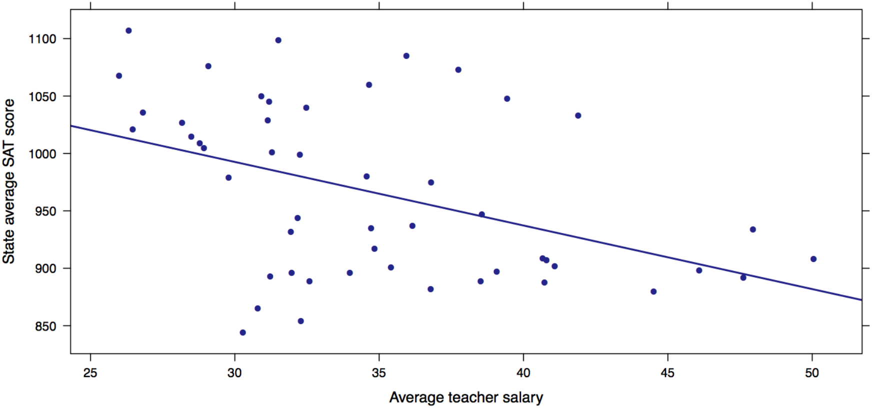
Potential paralysis?

- Wild: glimpses of a multivariate world
- Question: do we reinforce key aspects of *design* (observational data vs. randomized trials) when we teach inference?
- Do students infer that they can't make inferential conclusions if data don't arise from a randomized trial?

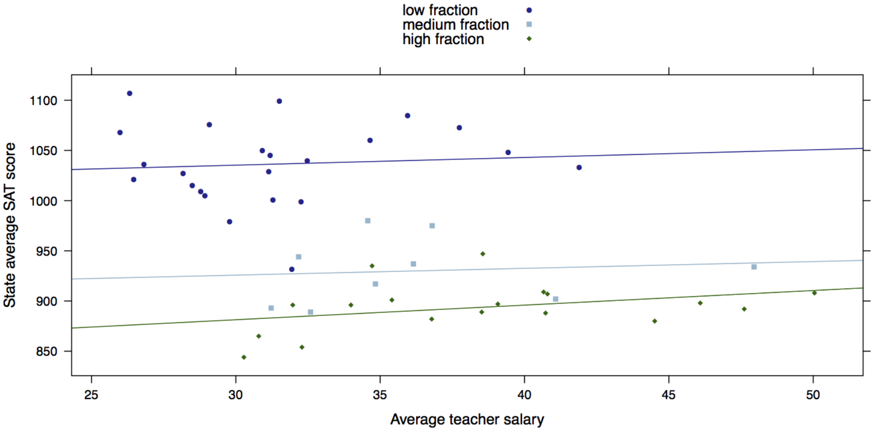
Potential paralysis?

- Wild: glimpses of a multivariate world
- Question: do we reinforce key aspects of *design* (observational data vs. randomized trials) when we teach inference?
- Do students infer that they can't make inferential conclusions if data don't arise from a randomized trial?
- What are implications in a world of found data?

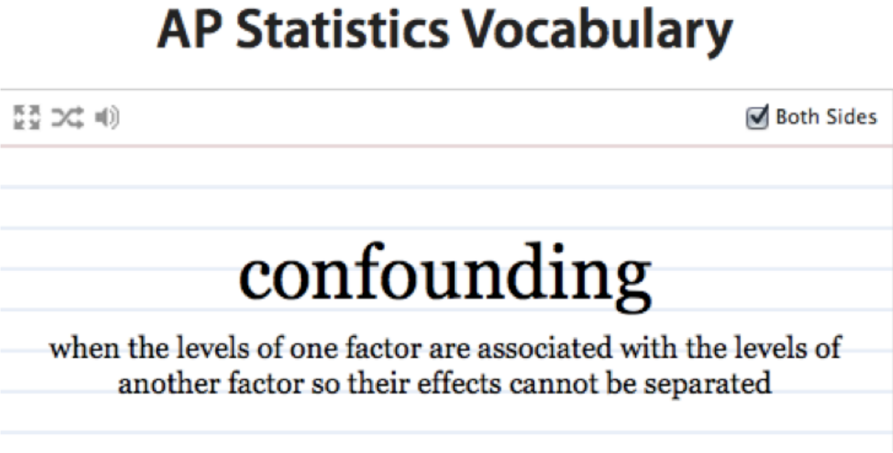
SAT scores and teacher salaries (state data from 2010)



SAT scores and teacher salaries (state data from 2010)



Multivariate thinking and confounding



AP Statistics Vocabulary

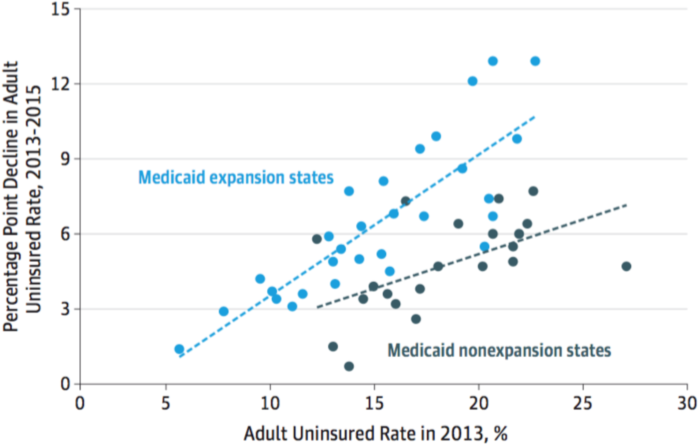
☑ Both Sides

confounding

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

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



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?)

(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

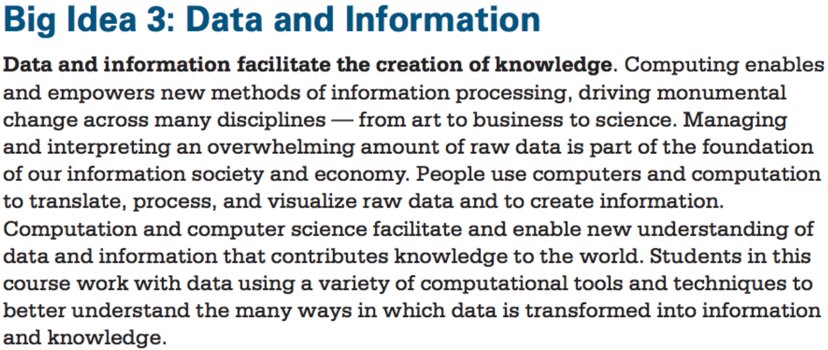
question: “what assumptions do you have students check when using the two sample t-test?”

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).

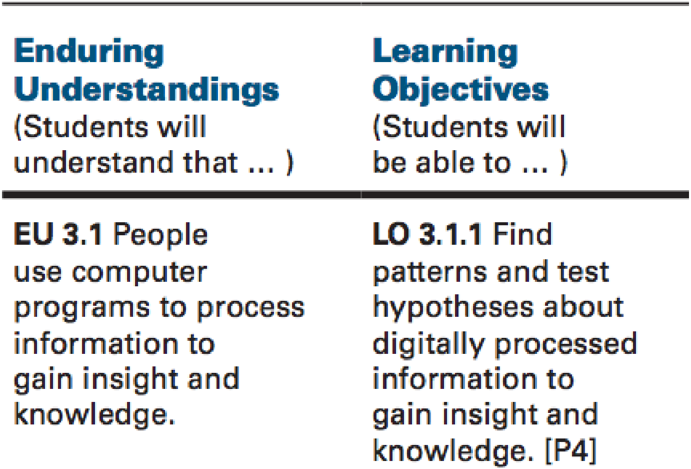
AP Computer Science Principles: an end-run around intro stat?



Big Idea 3: Data and Information

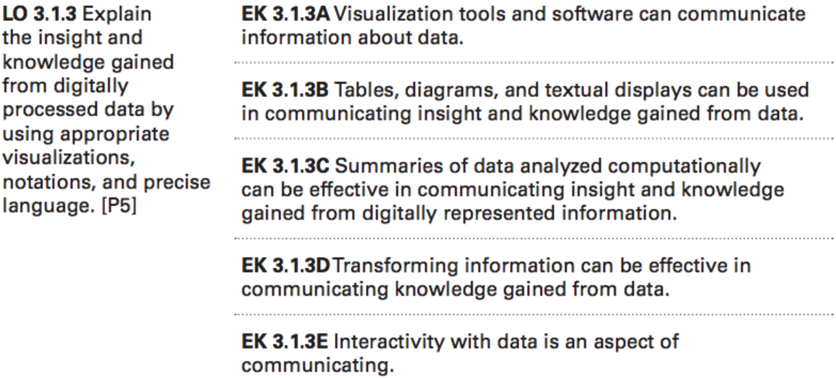
Data and information facilitate the creation of knowledge. Computing enables and empowers new methods of information processing, driving monumental change across many disciplines — from art to business to science. Managing and interpreting an overwhelming amount of raw data is part of the foundation 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 this year



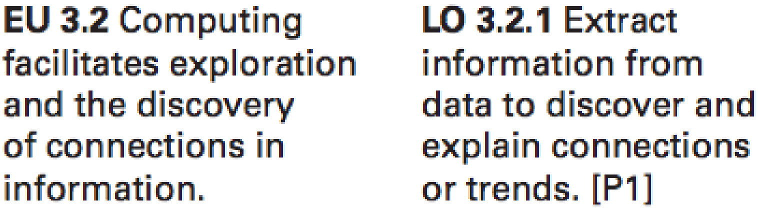
Enduring Understandings (Students will understand that ...)	**Learning Objectives** (Students will be able to ...)
EU 3.1 People use computer programs to process information to gain insight and knowledge.	**LO 3.1.1** Find patterns and test hypotheses about digitally processed information to gain insight and knowledge. [P4]

AP Computer Science Principles: project based learning



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]

Closing thoughts

- Teach (modern) design early and often
- Avoid paralysis: teach techniques to move beyond two-sample t-test (stratification and multiple regression)
- Make room by simplifying (what if all datasets were $n > 100$?)
- Show me the data: communicate the excitement of statistics as a way to extract meaning from data