Introductory Statistics in a World of Data Science: Where We Are and Where We Need to Head

Nicholas J. Horton

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

JSM IOL, July 29, 2018

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

Acknowledgements

- Dana Center Mathematics and Data Science Pathways Program (Rebecca Hartzler and Roxy Peck)
- Project MOSAIC: Danny Kaplan (Macalester College), Randy Pruim (Calvin College), and Ben Baumer (Smith College)
- Johanna Hardin (Pomona College) and the undergraduate guidelines working group
- revised GAISE College Report group
- those listed in the annotated bibliography in Table 2 of Horton and Hardin (TAS 2015)

Overview and motivation for the session

- Horton: Key changes needed for intro stat
- Nugent: Undergraduate data science and statistics programs
- Witten: Questions and challenges for the PhD curriculum in statistics

Slides and related resources at
`https://www.github.com/Amherst-Statistics/JSM2018`

Overview and motivation

Let students do what statisticians do: analyze non-trivial datasets by considering a variety of models, **using their imagination and developing their judgment** *in the process (Cobb, 1982).*

Overview and motivation

I believe it is the use of imagination and judgment that makes our subject appealing. **We owe it to our students not to keep that a secret** *(Cobb, 1982).*

NAS CATS (Committee of Applied and Theoretical Statistics) workshop

The growth that statistics has undergone is often not reflected in the education that future statisticians receive. There is a need to incorporate more meaningfully into the curriculum the **computational and graphical tools** *that are today so important to many professional statisticians. There is a need for improved training of statistics students in* **written and oral communication skills**, *which are crucial for effective interaction with scientists and policy makers.*

As academic statisticians, we are missing the boat. We are barking up the wrong tree. ... The kinds of statistics that we teach in undergraduate and especially in graduate programs have almost **nothing to contribute to anything that matters**. *... Then we wonder why the world passes us by.*

more from CATS workshop

The current curriculum in most statistics departments is, however, entirely too focused on hypothesis testing (Ed Rothman).

We risk being ignored if we do not stay relevant. (Carl Morris)

more from CATS workshop

The current curriculum in most statistics departments is, however, entirely too focused on hypothesis testing (Ed Rothman).

We risk being ignored if we do not stay relevant. (Carl Morris)

All are quotes from 1992.

ASA Undergrad Guidelines, 2014

We are concerned that many of our graduates do not have sufficient skills to be effective in the modern workforce. Thomas Lumley (personal communication) has stated that our students know how to deal with $n \to \infty$, but cannot deal with a million observations.

If statistics is the science of learning from data, then our students need to be able to "think with data" (as Diane Lambert of Google has so elegantly described).
- Horton and Hardin (TAS, 2015)

Where to Start? The Intro Stats Course

- the changing landscape of K-12 statistics education
- GAISE College Report (2016)
- what do we need to change?

Changing landscape of K-12 statistics education

Roxy Peck (JSM 2011) noted:

- statistics has been a recommended part of math curriculum for a long time
- recent developments have led to considerable more emphasis on statistics
- not just AP statistics: expectation for all students

Changing landscape of K-12 statistics education

1. Summarize, represent, and interpret data on a single count or measurement variable
2. Summarize, represent, and interpret data on two categorical and quantitative variables
3. Interpret linear models
4. Make inferences and justify conclusions from sample surveys, experiments, and observational studies

Inferential statistics

S-IC.3 Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC.4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

S-IC.5 Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.

S-IC.6 Evaluate reports based on data.

Changing landscape of K-12 statistics education

- Seven years later, where do things stand?
- SAT Math (`https://collegereadiness.collegeboard.org/sample-questions/math`)
- LOCUS (Levels of Conceptual Understanding in Statistics `https://locus.statisticseducation.org`)
- Eureka Math Curriculum (`https:greatminds.org`)

SAT Math test (example question)

A research assistant randomly selected 75 undergraduate students from the list of all students enrolled in the psychology-degree program at a large university. She asked each of the 75 students, "How many minutes per day do you typically spend reading?" The mean reading time in the sample was 89 minutes, and the margin of error for this estimate was 4.28 minutes. Another research assistant intends to replicate the survey and will attempt to get a smaller margin of error.

Which of the following samples will most likely result in a smaller margin of error for the estimated mean time students in the psychology-degree program read per day?

SAT Math test

A researcher wanted to know if there is an association between exercise and sleep for the population of 16-year-olds in the US. She obtained survey responses from a random sample of 2000 US 16-year-olds and found convincing evidence of a positive association between exercise and sleep. Which of the following conclusions is well supported by the data?

A: There is a positive association between exercise and sleep for 16-year-olds in the US

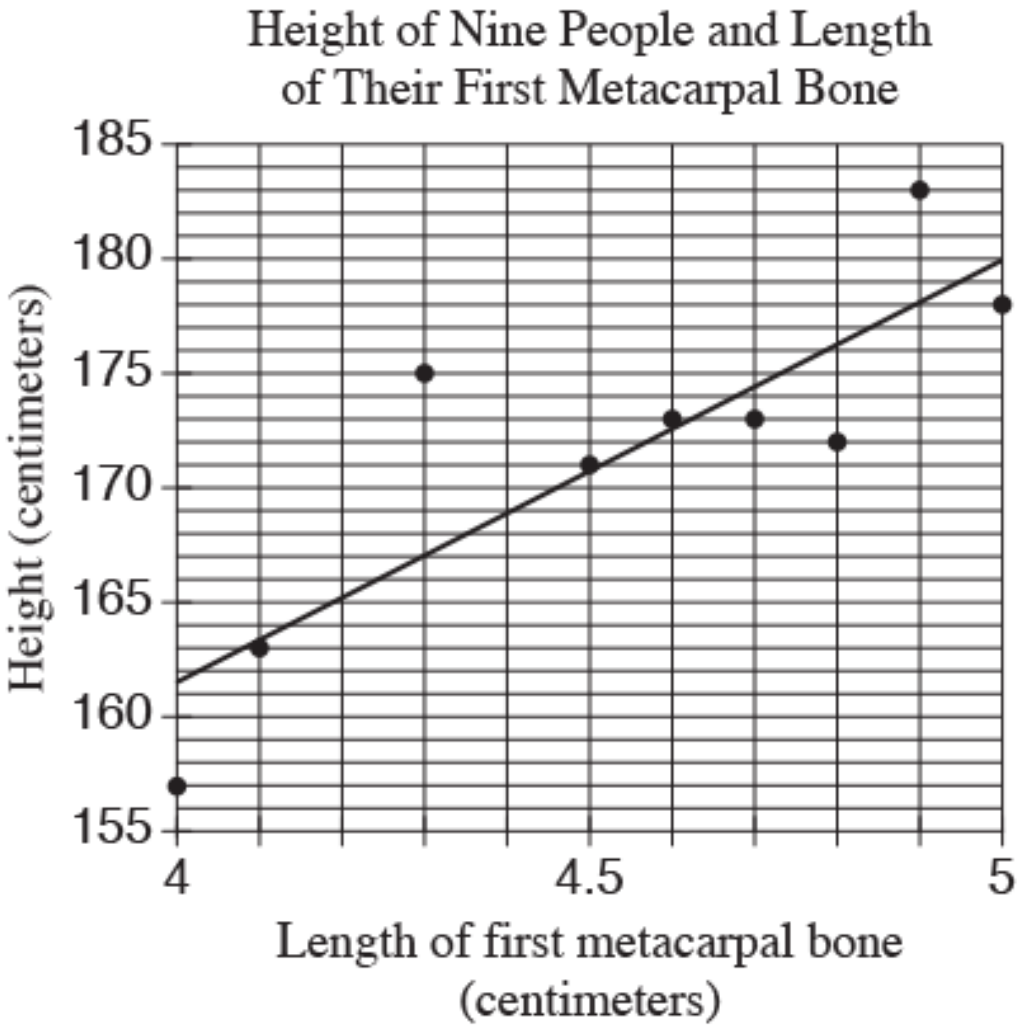
B: There is a positive association between exercise and sleep for 16-year-olds in the world.

C: Using exercise and sleep as defined by the study, an increase in sleep is caused by an increase of exercise for 16-year-olds in the US.

D: Using exercise and sleep as defined by the study, an increase in sleep is caused by an increase of exercise.

SAT Math: Linear models

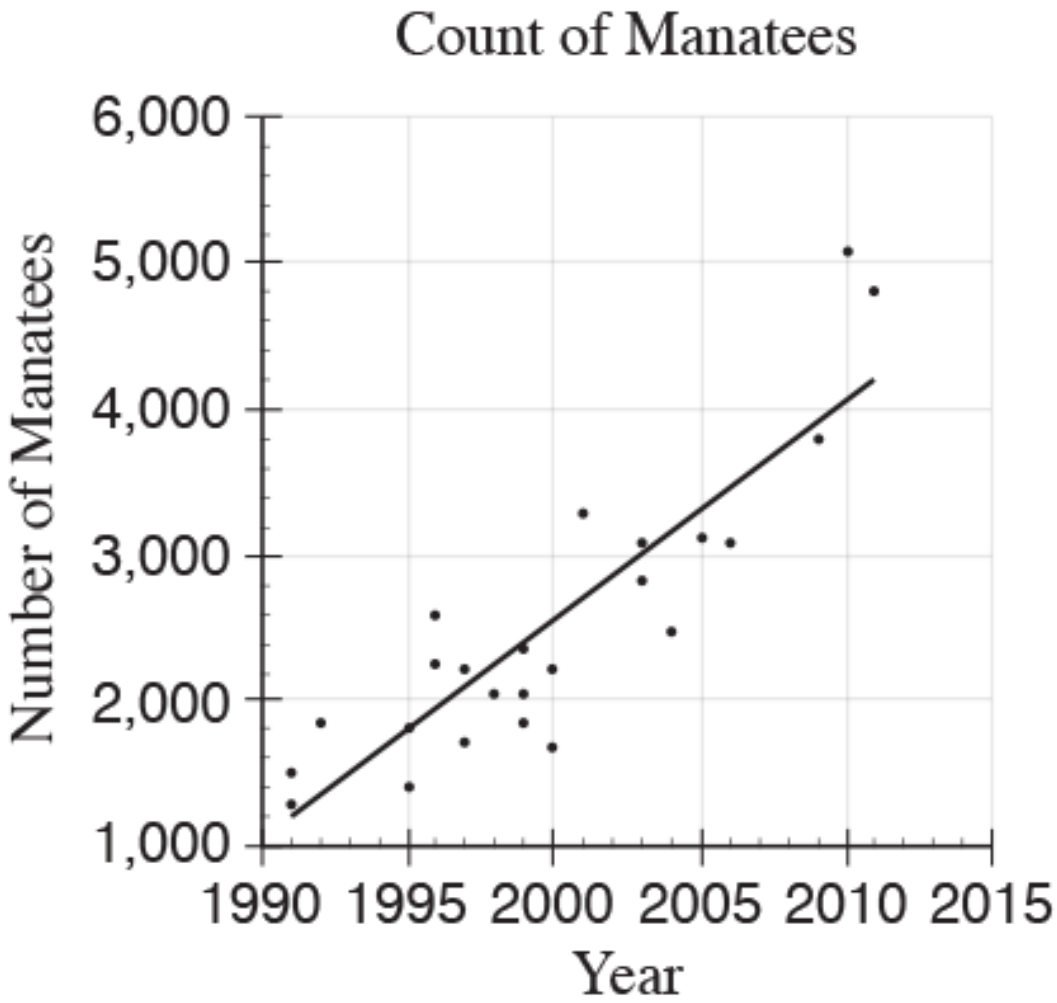
The first metacarpal bone is located in the wrist. The scatterplot below shows the relationship between the length of the first metacarpal bone and height for 9 people. The line of best fit is also shown.



SAT Math: Linear models

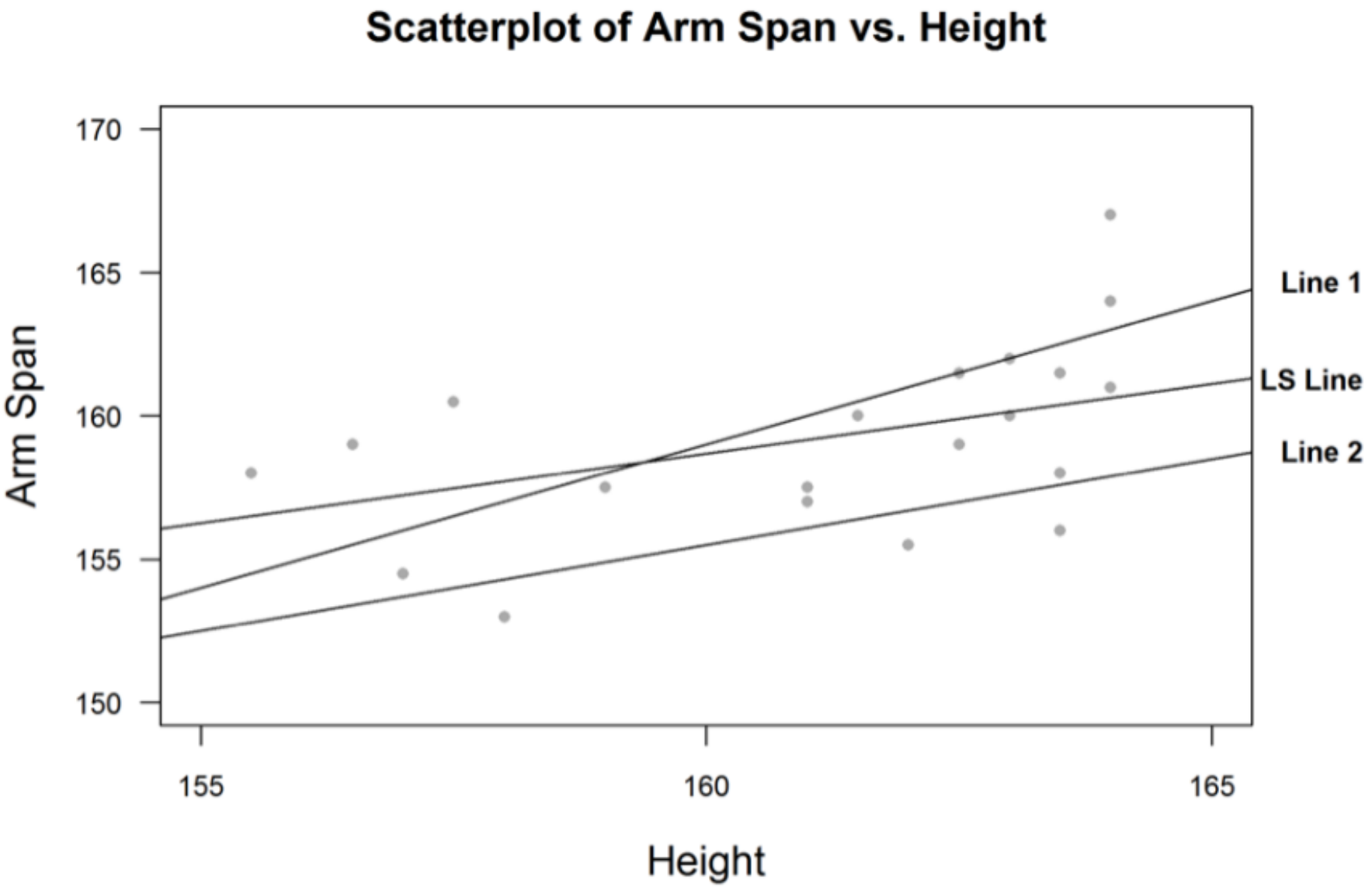
How many of the nine people have an actual height that differs by more than three centimeters from the height predicted by the line of best fit?

SAT Math: Linear models



The scatterplot above shows counts of Florida manatees, a type of sea mammal, from 1991 to 2011. Based on the line of best fit to the data shown, which of the following values is closest to the average yearly increase in the number of manatees?

LOCUS: Linear models



LOCUS: Linear models

The scatterplot below shows the relationship between height and arm span for a group of students. The least squares line (labeled LS line) and two other lines have been added to the scatterplot.

A: Compared to the other lines, Line 1 has the smallest sum of squared residuals.

B: The sum of squared residuals for Line 1 is greater than the sum of squared residuals for Line 2.

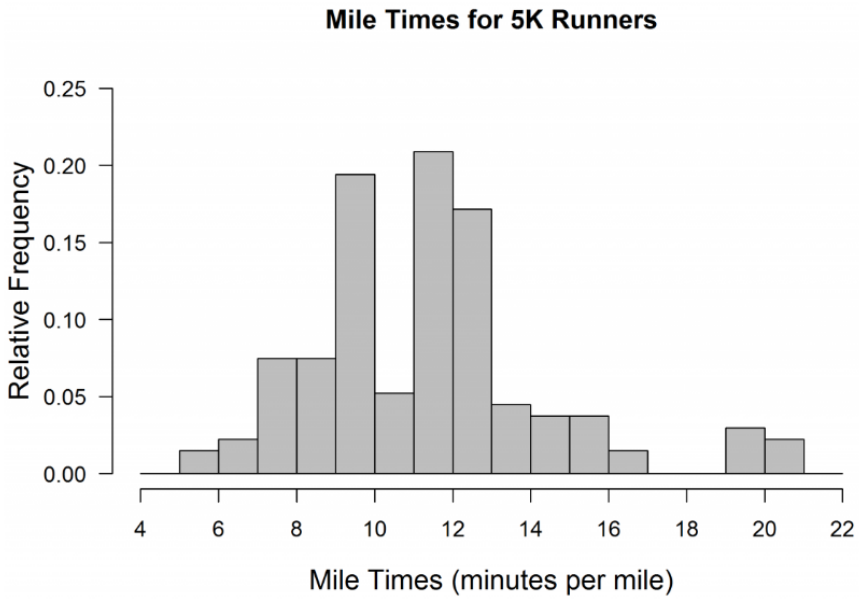
C: Compared to the other lines, the least squares line has the smallest sum of squared residuals.

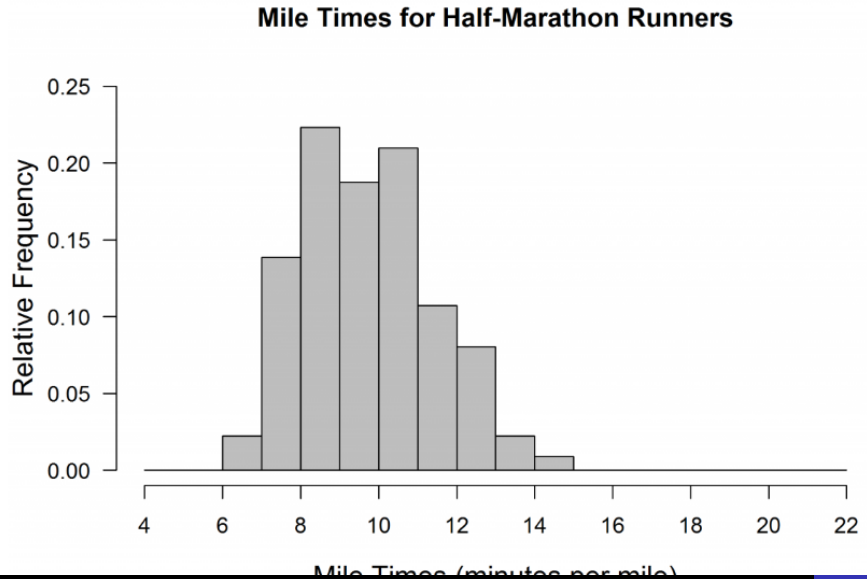
D: The sum of squared residuals for the least squares line is greater than the sum of squared residuals for Line 2.

LOCUS: Group differences

The city of Gainesville hosted two races last year on New Years Day. Individual runners chose to run either a 5K (3.1 miles) or a half-marathon (13.1 miles). One hundred thirty four people ran in the 5K, and 224 people ran the half-marathon. The mile time, which is the average amount of time it takes a runner to run a mile, was calculated for each runner by dividing the time it took the runner to finish the race by the length of the race. The histograms below show the distributions of mile times (in minutes per mile) for the runners in the two races.

LOCUS: Group differences





LOCUS: Group differences

Sierra predicted that, on average, the mile time for runners of the half-marathon would be greater than the mile time for runners of the 5K race. Do these data support Sierra's statement? Explain why or why not.

Eureka Math

NATIONAL IMPACT

Eureka Math is the most widely used math curriculum in the United States, according to a **study** released by the RAND Corporation. It is also the only curriculum found by **EdReports.org** to align fully with the Common Core State Standards for all grades, K–8. Additionally, over a dozen lessons from *Eureka Math* were rated to be EQuIP exemplars by **Achieve**.

Lesson Summary

When a single set of values is randomly divided into two groups:

- The two group means will tend to differ just by chance.
- The distribution of random groups' means will be centered at the single set's mean.
- The range of the distribution of the random groups' means will be smaller than the range of the data set.
- The shape of the distribution of the random groups' means will be symmetrical.

Exercises 5–9: Random Selection and Random Assignment

Take another look at the two studies described above. Study A (the dog food study) is an experiment, while study B (t messages) is an observational study. The term *random sample* implies that a sample was randomly selected from a population. The terms *random selection* and *random assignment* have very different meanings.

Random selection refers to randomly selecting a sample from a population. Random selection allows generalization to population and is used in well-designed observational studies. Sometimes, but not always, the subjects in an experime are randomly selected.

Random assignment refers to randomly assigning the subjects in an experiment to treatments. Random assignment allows for cause-and-effect conclusions and is used in well-designed experiments.

Lesson Summary

For a given sample, you can find the sample mean.

- There is variability in the sample mean. The value of the sample mean varies from one random sample to another.
- A graph of the distribution of sample means from many different random samples is a simulated sampling distribution.
- Sample means from random samples tend to cluster around the value of the population mean. That is, the simulated sampling distribution of the sample mean will be centered close to the value of the population mean.
- The variability in the sample mean decreases as the sample size increases.
- Most sample means are within two standard deviations of the mean of the simulated sampling distribution.

Changing landscape of K-12 statistics education

Executive Summary: Roxy Peck was right!

- intro stats will not be the first exposure for main topics
- this is good, since GAISE K-12 report talks about the need for repeated exposure
- some material may be reviewed more quickly

Changing landscape of K-12 statistics education

Caveats:

- not all students will see all of this material
- not all teachers are prepared to teach this material
- huge disparities exist in our K-12 system

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

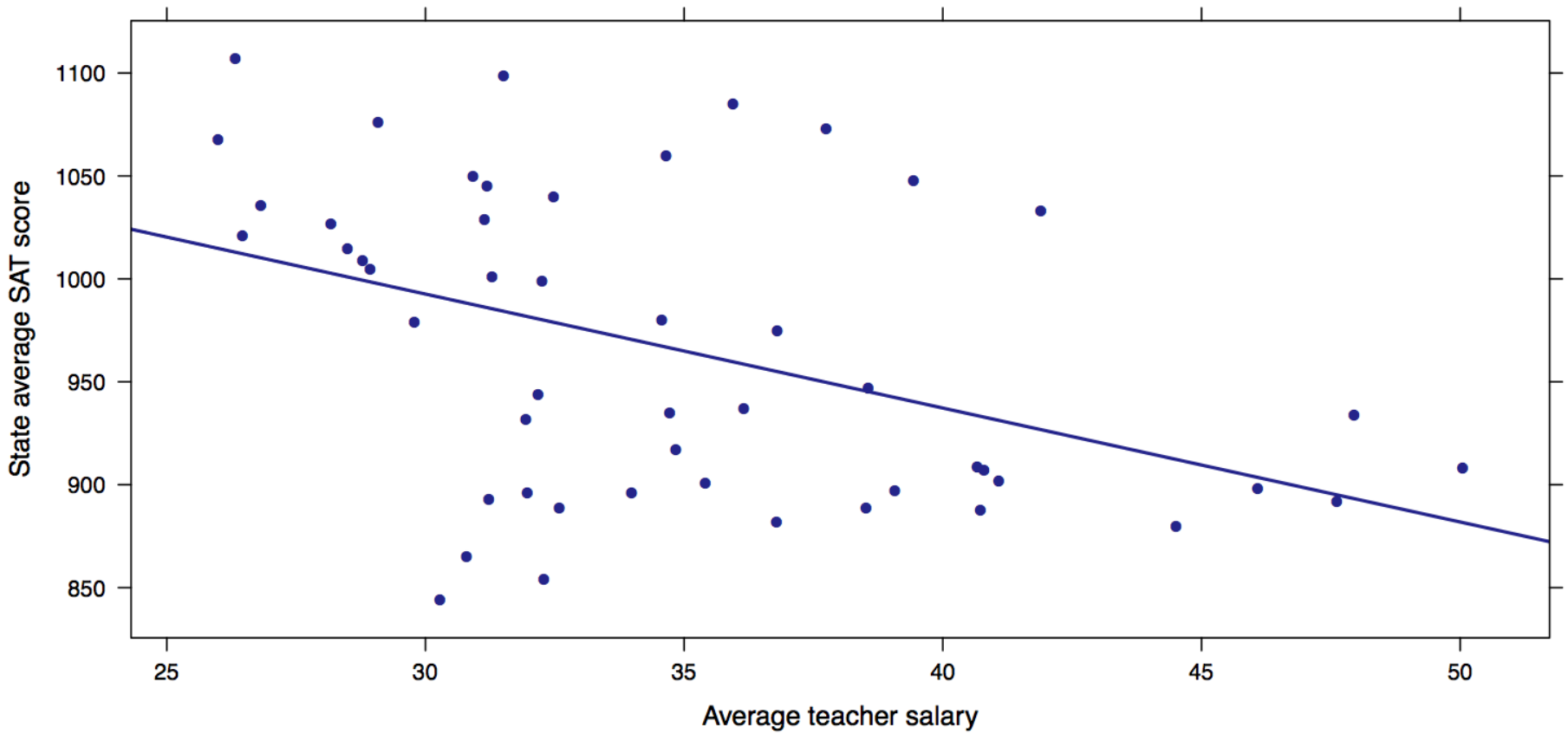
revised GAISE College Report (2016)

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.

Some big ideas to bring into intro stats

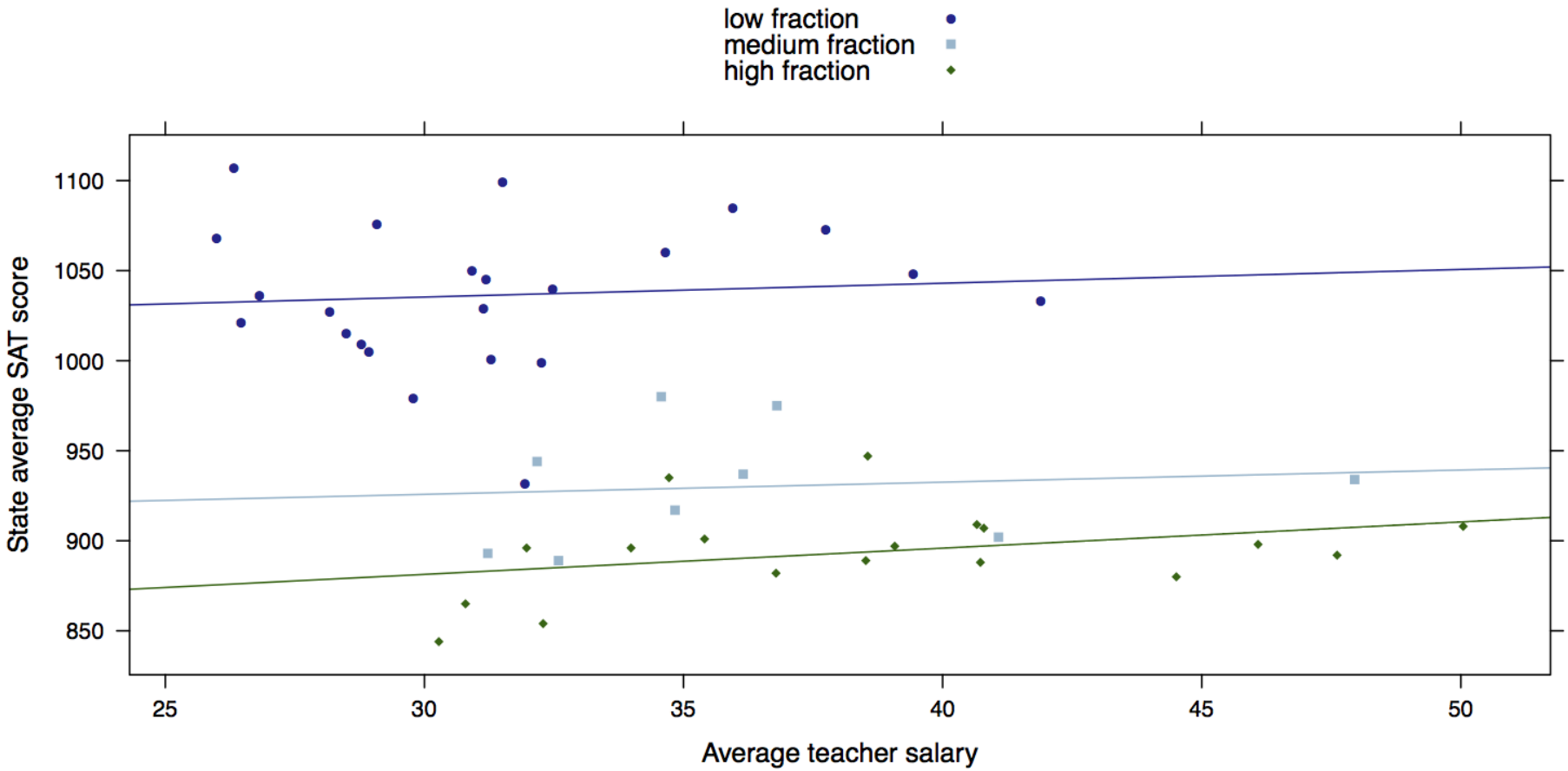
1. Give students experience with **multivariable thinking**
2. Use **technology to explore concepts and analyze data**

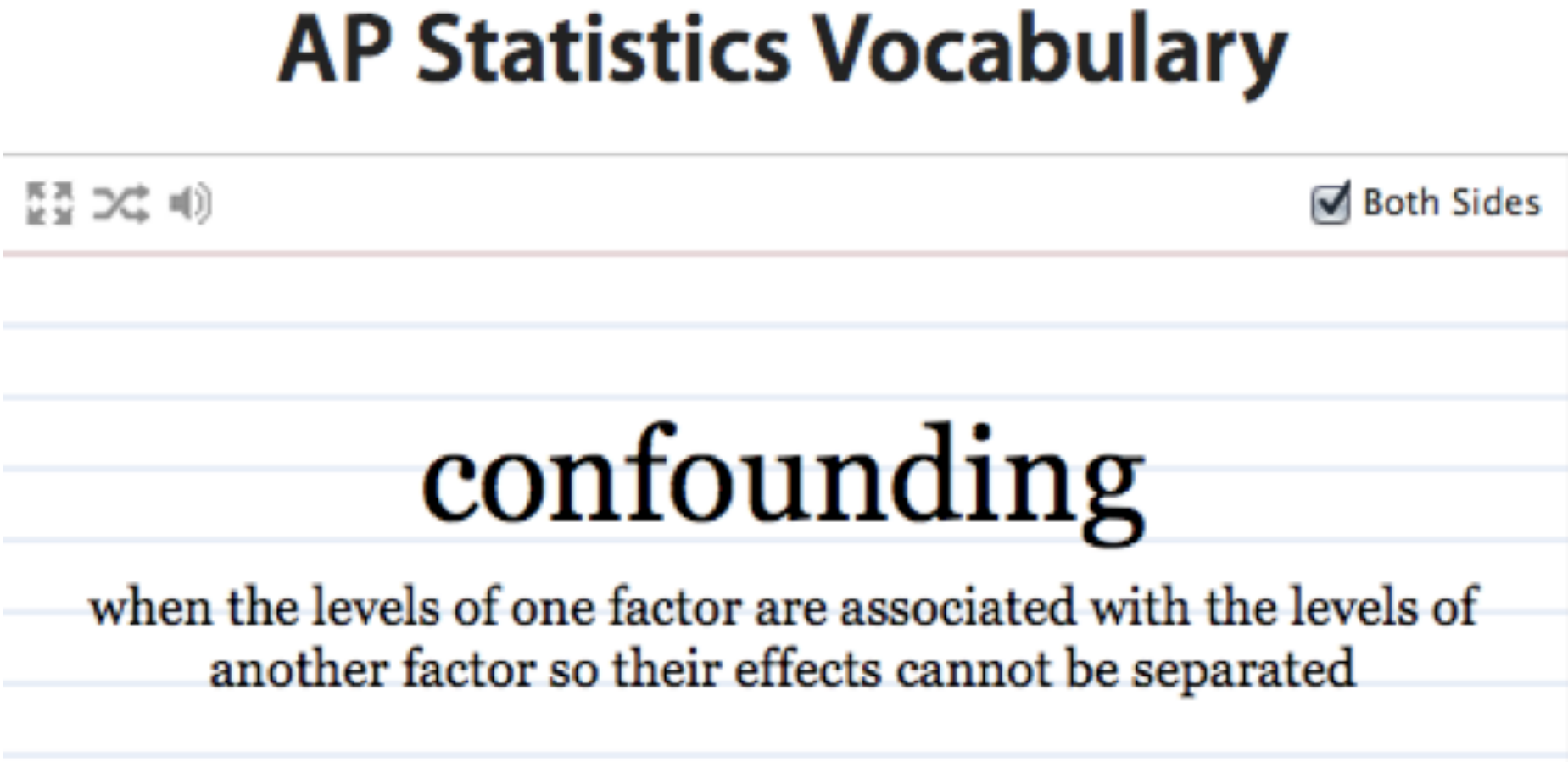
Multivariate thinking and confounding



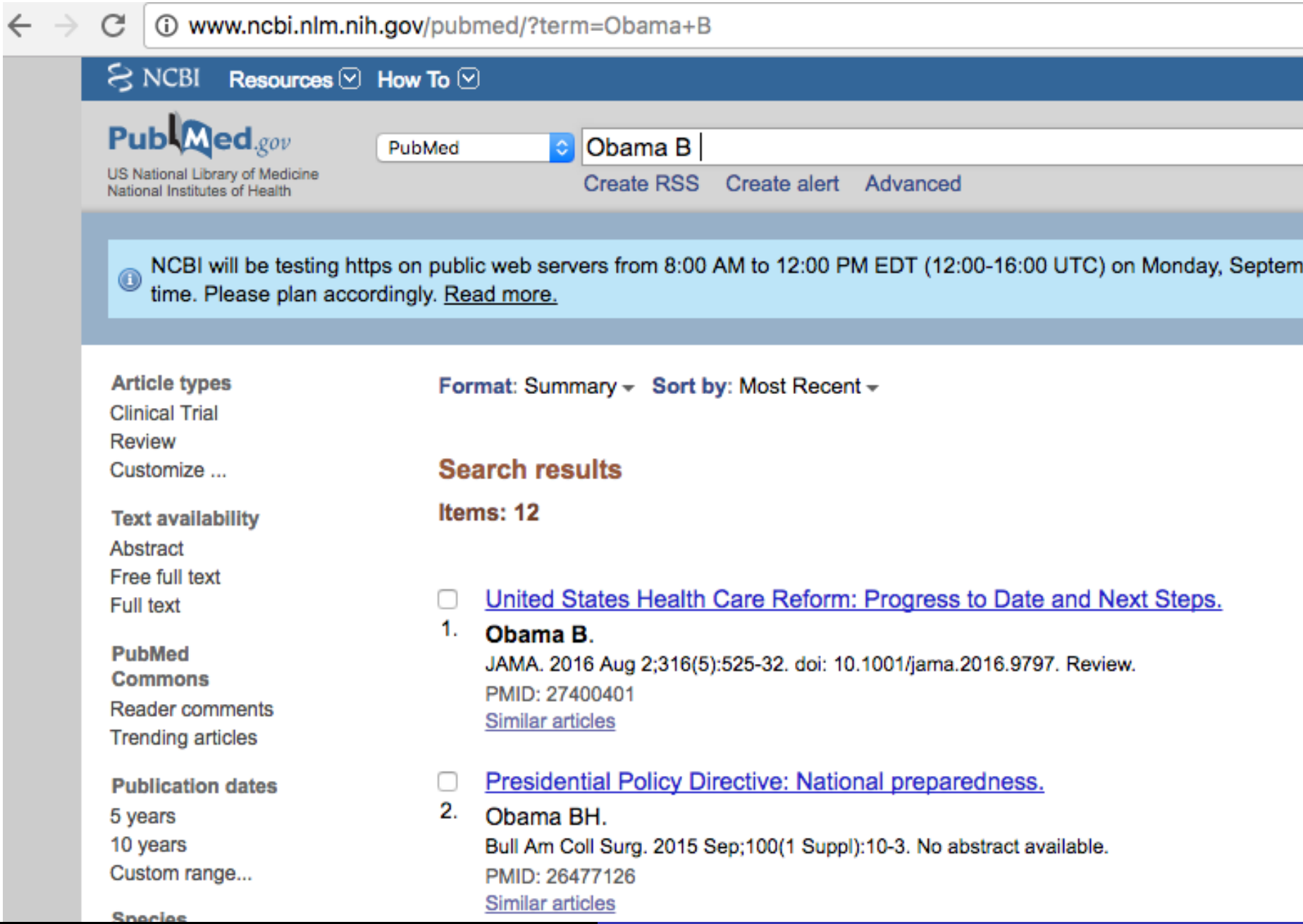
College entrance scores and teacher salaries (US state data from 2010)

Multivariate thinking and confounding



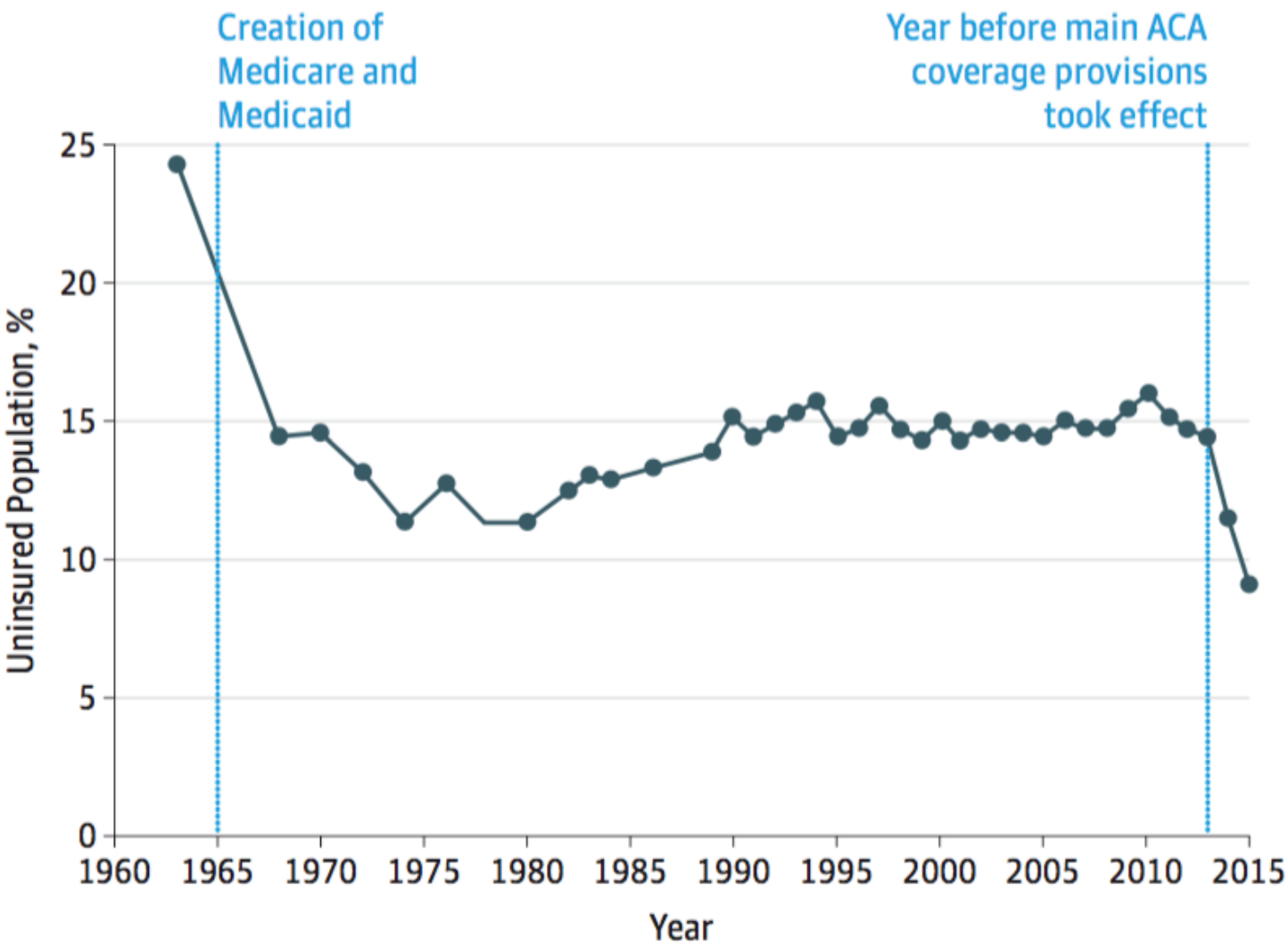

AP Statistics Vocabulary
Both Sides
confounding
when the levels of one factor are associated with the levels of another factor so their effects cannot be separated

Design and confounding: President Obama's publications



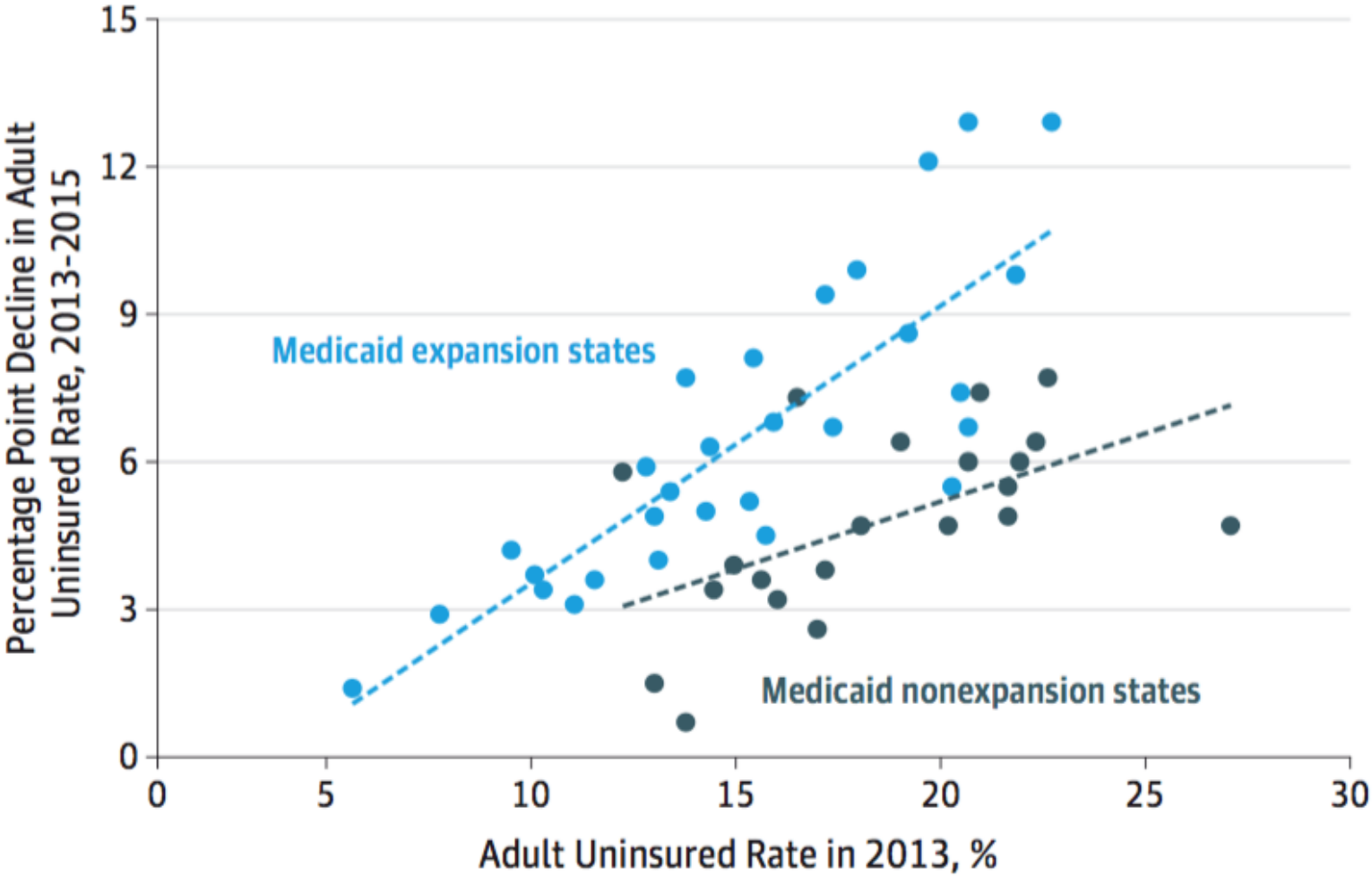
Obama's single author JAMA paper

Figure 1. Percentage of Individuals in the United States Without Health Insurance, 1963-2015



Obama's single author JAMA paper

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



Kari Lock talk (Tuesday at 2:00pm)

Multivariable Thinking with Data Visualization

`https://www.github.com/Amherst-Statistics/JSM2018`

Teaching with R: one of several solutions

The New York Times

Business Computing

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Data Analysts Captivated by R's Power

Left, Stuart Isett for The New York Times; right, Kieran Scott for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

FACEBOOK
TWITTER

Need to be disciplined and keep it simple

An analyst wants to calculate the mean `pH` of assays from two treatments. What's the simplest way to do this in base R? Using other packages?

Possible answer

```
> with(chem, aggregate(pH, by=list(treat),
  FUN=mean, na.rm=TRUE, simplify=TRUE))
  Group.1        x
1    grpA 1.904762
2    grpB 1.756757
```

Possible answer

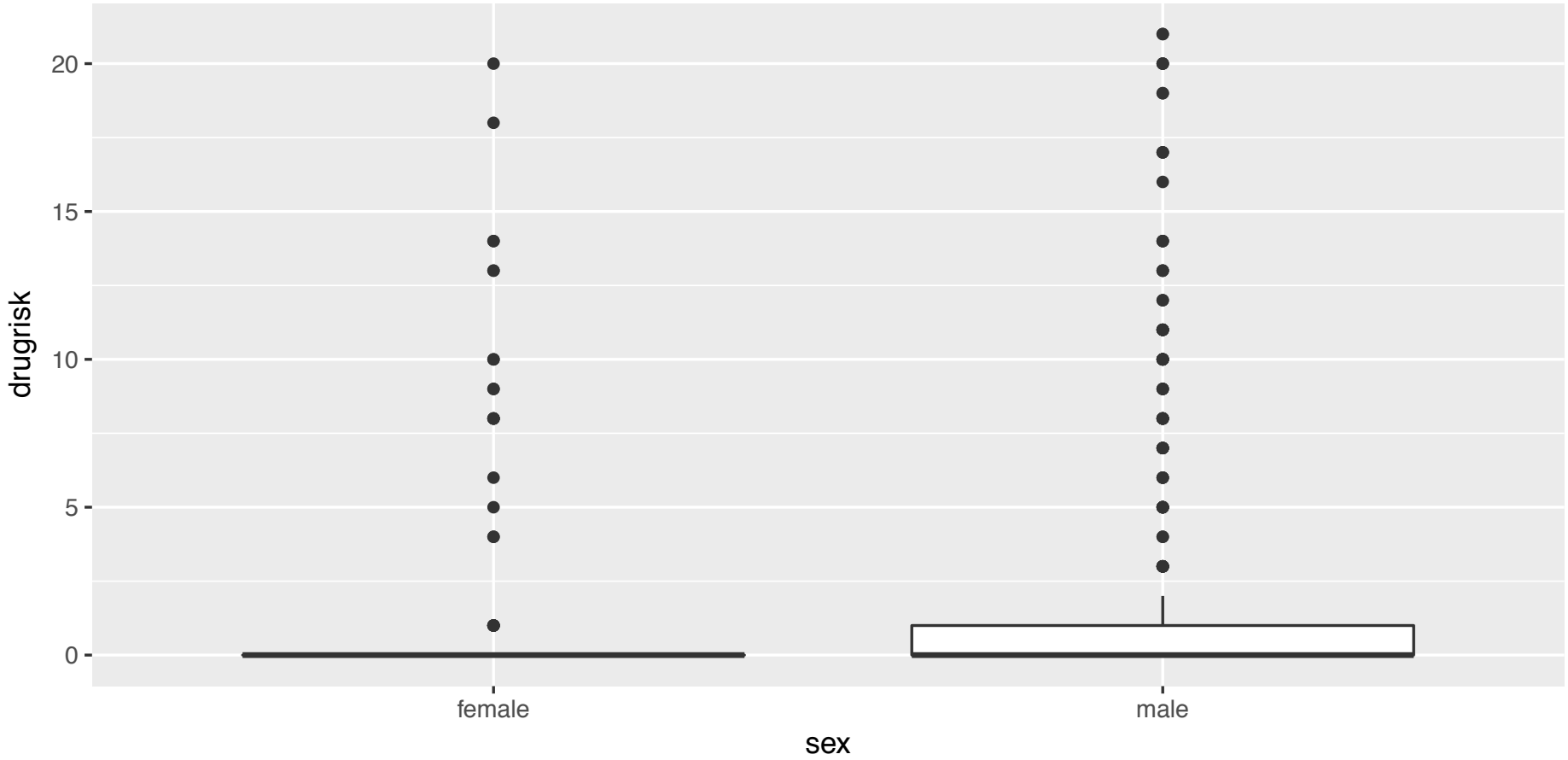
```
> with(chem, tapply(pH, treat, mean, na.rm=TRUE))
    grpA     grpB
1.904762 1.756757
```

A better answer (suitable for intro)

```
> library(mosaic)
> favstats(pH ~ treat, data = chem)
     sex min Q1 median Q3 max mean   sd   n missing
1   grpA   0  0      0  1  11 1.90 4.37 357       2
2   grpB   0  0      0  0  10 1.76 4.15 111       0
```

mosaic modeling language (Y ~ X)

```
> gf_boxplot(pH ~ treat, data = chem)
```



mosaic modeling language (Y ~ X)

```
> lm(pH ~ treat, data = chem)

Coefficients:
(Intercept)          grpB
      1.905        -0.148
```

One simple approach to:

- generate descriptive statistics
- create graphical displays
- fit regression models

See *R Journal* paper (journal.r-project.org/archive/2017/RJ-2017-024) and Little Books (www.github.com/ProjectMOSAIC/LittleBooks)

Are these just ‘training wheels‘?

- builds on the ‘formula‘ object in R
- extends to more sophisticated models
- uses a consistent and coherent syntax
- utilizes tidyverse idioms for data wrangling
- should we teach ggplot2 to beginners? (`http://varianceexplained.org/r/teach-tidyverse`)

Simplified access using cloud computing

- "bring a browser model" for RStudio and JupyterHub
- minimize cognitive load early on during a course
- avoid all complications of software and package installation
- let students do interesting things on day one (Wang et al, TISE, 2017, https://escholarship.org/uc/item/84v3774z)
- pedagogy and technology are tightly linked (Cetinkaya-Rundel and Rundel, TAS, 2018)
- costs of software have gone to zero
- costs of cloud servers have gone down dramatically
- configuration and installation have gotten much simpler (but need IT to assist, not get in the way)

Communication and knowledge transference

ASA undergrad guidelines for statistics programs:

- Students need to be able to **communicate** complex statistical methods in basic terms to managers and other audiences and to visualize results in an accessible manner.
- They must have a clear understanding of **ethical standards**.
- Programs should provide **multiple opportunities to practice and refine** these statistical practice skills and use of analysis cycle.

R Markdown and reproducible analysis

The ability to express statistical computations is an essential skill (Nolan and Temple Lang, TAS 2010)

- R Markdown used as first workflow for introductory statistics students at colleges and universities all over the country (Baumer et al, *TISE*, 2014)
- easy to deploy as a cloud application (fewer barriers for students)
- forms a 'necessary but not sufficient' component of reproducible research
- tightly integrated into RStudio (designed for experts, useful for newbies)
- also available in environments such as Jupyterhub

Challenges and opportunities for statistics

Key learning outcomes from Berkeley's Data8.org course

- Calculate specified statistics of a given dataset.
- Identify the sources of randomness in an experiment.
- Correctly generate and interpret histograms, bar charts, and box plots.
- Formulate a null hypothesis that relates to a given question, which can be assessed using a statistical test.
- Carry out statistical analyses including computing confidence intervals and performing hypothesis tests in a variety of data settings.

Key learning outcomes from Berkeley's Data8.org course (cont.)

- Given the result of a statistical analysis from the course, form correct conclusions about a question based on its meaning.
- Given a question and an analysis, explain whether the analysis addresses the question and how the analysis could change and still address the question.
- **Correctly make predictions using regression and classification techniques.**
- **Assess the accuracy and variability of a prediction.**
- (Plus how to write a function!)

Challenges and opportunities

Read the Introducing AP Computer Science Principles video transcript

Computer Science: The New Literacy

Whether it's 3-D animation, engineering, music, app development, medicine, visual design, robotics, or political analysis, computer science is the engine that powers the technology, productivity, and innovation that drive the world. Computer science experience has become an imperative for today's students and the workforce of tomorrow.

The AP Program designed AP Computer Science Principles with the goal of creating leaders in computer science fields and attracting and engaging those who are traditionally underrepresented with essential computing tools and multidisciplinary opportunities.

Largest first year AP exam ever in 2017 (45,000 students took the exam)

Challenges and opportunities

Read the Introducing AP Computer Science Principles video transcript

Computer Science: The New Literacy

Whether it's 3-D animation, engineering, music, app development, medicine, visual design, robotics, or political analysis, computer science is the engine that powers the technology, productivity, and innovation that drive the world. Computer science experience has become an imperative for today's students and the workforce of tomorrow.

The AP Program designed AP Computer Science Principles with the goal of creating leaders in computer science fields and attracting and engaging those who are traditionally underrepresented with essential computing tools and multidisciplinary opportunities.

Largest first year AP exam ever in 2017 (45,000 students took the exam)
Second year (numbers still rough) more than 83,000 students took the exam

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.

Challenges and opportunities

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]

Challenges and opportunities

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.

Challenges and opportunities

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]

Should all statistics students be programmers?

No!

July 2018

Hadley Wickham
@hadleywickham
Chief Scientist, **RStudio**

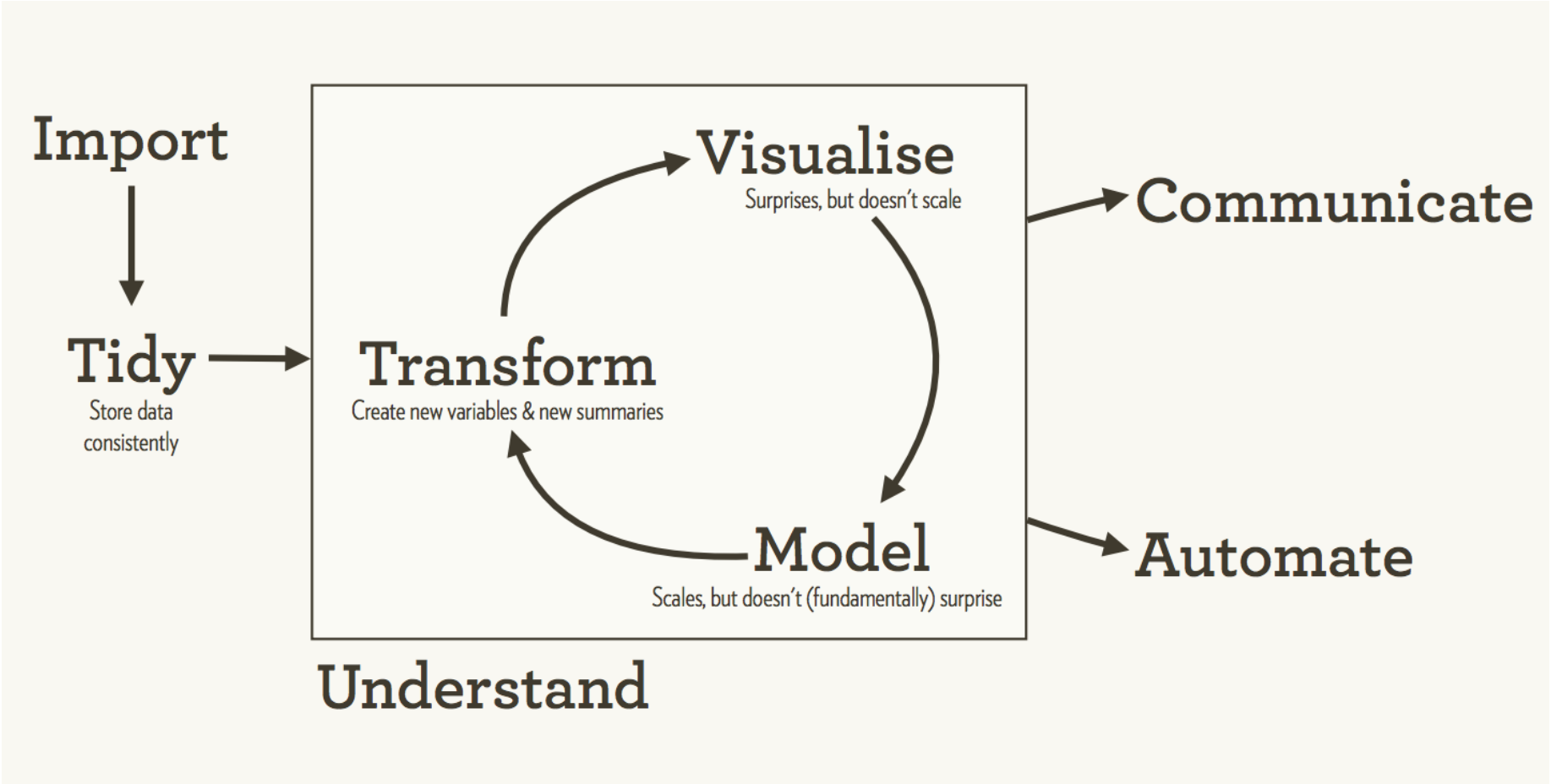
Should all statistics students program?

July 2018

Yes!

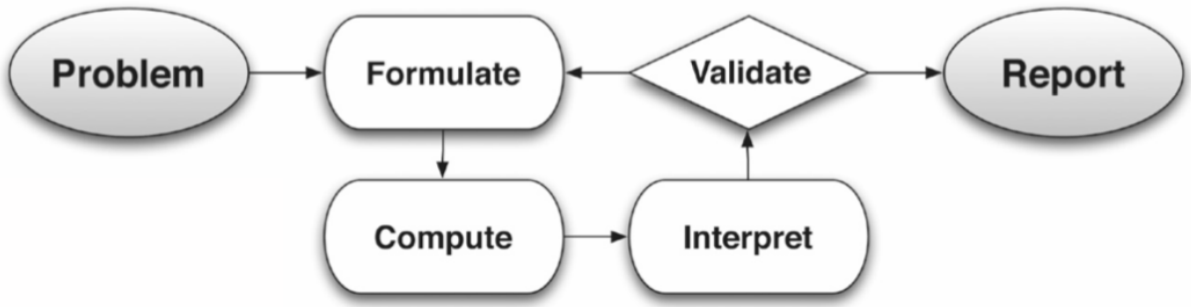
Hadley Wickham
@hadleywickham
Chief Scientist, **RStudio**

Data analysis cycle



OVERVIEW

In Grade 8, students used functions for the first time to construct a function that models a linear relationship between two quantities (**8.F.B.4**) and to describe qualitatively the functional relationship between two quantities by analyzing a graph (**8.F.B.5**). In the first four modules of Algebra I, students learn to create and apply linear, quadratic, and exponential functions in addition to square and cube root functions (**F-IF.C.7**). In Module 5, they synthesize what they have learned during the year by selecting the correct function type in a series of modeling problems without the benefit of a module or lesson title that includes function type to guide them in their choices. This supports the CCLS requirement that student's use the modeling cycle, in the beginning of which they must formulate a strategy. Skills and knowledge from the previous modules support the requirements of this module, including writing, rewriting, comparing, and graphing functions (**F-IF.C.7**, **F-IF.C.8**, **F-IF.C.9**) and interpretation of the parameters of an equation (**F-LE.B.5**). Students also draw on their study of statistics in Module 2, using graphs and functions to model a context presented with data and tables of values (**S-ID.B.6**). In this module, we use the modeling cycle (see page 72 of the CCLS) as the organizing structure rather than function type.



Closing thoughts

- lots of changes at the K-12 level make it possible for us to rethink focus of intro stat
- caveat: students will need refreshers and additional practice to improve conceptual understanding
- improved tools have make it easier to extract meaning from data
- era of (cheap) cloud computing: transformative opportunities to simplify access for students

Closing thoughts

- add more multivariate thinking (and multiple regression) in intro stats (MLR now 20% of our intro course)
- use project-based learning to teach statistics and data science analysis cycle and reproducible workflows
- de-emphasize p-values (plug for Allen Downey's "Inference in Three Hours, and More Time for the Good Stuff", `https://www.github.com/Amherst-Statistics/JSM2018`)
- spend time exploring YOUR courses and support faculty working to update them

Introductory Statistics in a World of Data Science: Where We Are and Where We Need to Head

Nicholas J. Horton

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

JSM IOL, July 29, 2018

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