Data science is going to change the way we teach statistics

Matthew Brett

Source

https://github.com/matthew-brett/psy-data-sci

Plan

- old and new;
- new is heading our way;
- data science by example;
- why it matters;
- how it can help us teach statistics.

Old and new

The perils of the old.

A tipping point

"Mere renovation is too late: we need to rethink our undergraduate curriculum from the ground up" – George W Cobb (2015).

https://nhorton.people.amherst.edu/mererenovation

A tipping point

- http://data8.org
- “The course is designed for entry-level students from any major. It is designed specifically for students who have not previously taken statistics or computer science courses.” (link).
- Python and data science
- R.

Defining data science

https://github.com/hadley/stats337/blob/master/README.md

Data science by example

https://github.com/matthew-brett/psy-data-sci/brexit.ipynb

Themes

- code at the heart of analysis;
- general programming language as foundation;
- real data, real questions;
- flexible;
- reproducible;
- documented.

The efficiency gap

Excel: my part in its downfall

- 1986 - programming optional;
- 1996 - an uneasy truce;
- 2006 - the rise of R and Python;
- 2016 - Jupyter, Pandas, RStudio.

Can we really teach code?

- understand what algorithms are, how they are implemented as programs on digital devices, and that programs execute by following precise and unambiguous instructions;
- create and debug simple programs;
- use logical reasoning to predict the behaviour of simple programs
- use technology purposefully to create, organise, store, manipulate and retrieve digital content

What do we gain?

- change in culture;
- technical gains;
- deeper insight into algorithms.

Change in culture

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. — John W. Tukey (1962) "The Future of Data Analysis".

Change in culture

> *a) Focus on finding a good solution - that's what consultants get paid for.*
> *b) Live with the data before you plunge into modeling.*
> *c) Search for a model that gives a good solution, either algorithmic or data.*
> *d) Predictive accuracy on test sets is the criterion for how good the model is.*
> *e) Computers are an indispensable partner*
> *— Leo Breiman (2001) "Statistical modeling: the two cultures"*

A change in culture

Computing results are now being presented in a very loose, "breezy" way—in journal articles, in conferences, and in books. All too often one simply takes computations at face value. This is spectacularly against the evidence of my own experience. I would much rather that at talks and in referee reports, the possibility of such error were seriously examined.
– David L. Donoho (2010). An invitation to reproducible computational research. Biostatistics Volume 11, Issue 3 Pp. 385-388

Technical gains

- flexibility in data that can be analyzed (Twitter feeds, web pages, databases, geospatial data. . .);
- flexibility in data exploration, cleaning, visualization;
- flexibility in analysis methods;
- large active online community

Deeper insight with algorithms

Statistical reasoning is much easier to understand in terms of algorithms and computation instead of mathematics (Cobb 2015; Hesterberg 2015). Courses should teach “computational thinking” (Temple Lang 2015);

Resampling statistics

Julian Simon (1997) “Resampling, the new statistics” at http://www.resample.com/intro-text-online.

Trials of teaching method in high school students (1969) and undergraduates (1976) including those with “low skills and little interest in mathematics”.

http://www.jstor.org/stable/pdf/27958125.pdf

Opening the black box

“What I cannot create, I do not understand”

Found on Richard Feynman’s blackboard after his death.

A short course on statistics with almost no mathematics

- permutation test;
- correlation by permutation;
- traditional ANOVA;
- Maybe permutation ANOVA;

A prediction

- in 10 years, all traditional “statistics” courses will be data science courses;
- they will be called something like “data science” or “data analysis”;
- our students will be much more effective as a result.

Is this the end?

Yes, it's the end of the talk.

References

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3). Institute of Mathematical Statistics: 199–231. https://projecteuclid.org/euclid.ss/1009213726.

Cobb, George. 2015. "Mere Renovation Is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground up." *The American Statistician* 69 (4). Taylor & Francis: 266–82.

Hesterberg, Tim C. 2015. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." *The American Statistician* 69 (4). Taylor & Francis: 371–86.

Simon, Julian L, and Allen Holmes. 1969. "A New Way to Teach Probability Statistics." *The Mathematics Teacher* 62 (4). JSTOR: 283–88.

Simon, Julian L, David T Atkinson, and Carolyn Shevokas. 1976. "Probability and Statistics: Experimental Results of a Radically Different Teaching Method." *The American Mathematical Monthly* 83 (9). JSTOR: 733–39.