Moving Away from Ad Hoc Statistical Computing Education

Colin Rundel
Duke University
JSM 2017 - Baltimore
Claim:

Computation is fundamental to (modern) statistics, but we often don’t teach it that way.
Where we started:
An initial foray into teaching computation ...

- OpenIntro R labs created in Fall 2011
  - Used by two Intro Stat courses (Sta 10 & Sta 101) that year,
  - and Sta 102 the next Fall.

- Right place at the right time for R,
  - RStudio server beta (Fall 2011)
  - knitr & R Markdown (Summer 2012)
Reproducible computation

- Audience are non-stat majors coming from social and life sciences
- Teach tools and methods that can grow with the students
- Teach workflows that are inherently reproducible

Scriptability + Literate Programming + Version Control

RStudio
Where we were:
How are we doing?
Statistical Computing (Sta 323)

- Optional 2nd / 3rd year elective with minimal prereqs
- Offered once a year in the Spring, ~30-40 students
- Focus on team based open-ended problem solving

A programming course with statistics vs. A statistics course with programming
Computational Skills in Sta 323

- Programming (R)
- Reproducibility (R
- Version control (GitHub
- Unix shell (bash)
- Data munging (dplyr)
- Web scraping (scrapy)

- Web APIs (JSON, XML)
- Visualization (ggplot2)
- Databases and SQL (SQL)
- Distributed computing (Spark, Sparklyr)
- Interactivity (Shiny)
- ...

...
Rethinking gateways

• Students are desperate to learn “Data Science”

• Identify the skills students find useful / exciting while still preparing students for later requirements (theoretical and computational)
  • Teach the *good* stuff in R (visualization, munging, scraping, etc.)
  • Computation complements theory

Where we are:

1st Year
- Intro Stat
- Intro Data Science

2nd Year
- Regression
- Math Stat

3rd Year
- Bayesian
- Statistical Computing
- Other Electives

4th Year
- Case Studies
- Thesis

Options:
- CS 101 or 201
- or
- Other
Lessons Learned / Future Thoughts

• There was never a master plan, mostly an exercise in identifying what we could do better / filling in the gaps

• Immediate value + downstream effects

• Healthy cross-pollination between our undergraduate and master’s curricula - demand for the same shared core skills

• Staying current is hard, but immensely valuable

• Coordination between courses is still a work in progress
Questions / Comments?

rundel@gmail.com

@rundel

github.com/rundel

github.com/rundel/Presentations/

- Sta 101 - Data Analysis and Statistical Inference
- Sta 102 - Intro Biostatistics
- Sta 112 - Better Living through data science
- Sta 323 - Statistical computing