Expectations and Skills for Undergraduate Students Doing Research in Statistics and Data Science

Jo Hardin
Pomona College

jo.hardin@pomona.edu
@jo_hardin47
Github: hardin47
High-Impact Educational Practices

A Brief Overview


Chart of High-Impact Practices (pdf)

High-Impact Educational Practices: A Brief Overview

The following teaching and learning practices have been widely tested and have been shown to be beneficial for college students from many backgrounds. These practices take many different forms, depending on learner characteristics and on institutional priorities and contexts.
Needed skills for successful undergrad research

- Make an argument
- Engage with theory
- Work independently
- Wrangle data
My background

- Research with ~4 students / summer
- Senior thesis projects ~4 students / year
- 12 peer-reviewed pubs w/undergrads, 1 submitted
Make an Argument

- Theoretical
- Simulation
- Literature

In the Classroom

- Hypothesis test using mean and median to demonstrate two different approaches to the same scientific hypothesis
- Simulation studies (e.g., bootstrap confidence intervals)
- How do we know that? How can we argue that result is better than the other?
Make an Argument

- As compared to what?
- Permutation test for cutoff
- Simulation analysis to demonstrate FP & FN rates


Resistant multiple sparse canonical correlation.

Coleman J, Repogle J, Chandler G, Hardin J.
Engage with Theory

- Connect theoretical ideas to core principles in statistics

In the Classroom

- Moment generating function: why do they uniquely determine a distribution?

- Simulate theoretical results for visualization of the process.
Engage with Theory

- June 2016 – July 2017
- Kept coming back to his knowledge of prediction intervals for SLR (MSE)
- Organization was impeccable

Constructing Prediction Intervals for Random Forests

**Author:**
Benjamin Lu

**Advisor:**
Dr. Jo Hardin
Work Independently

- What did you do?
- Why did you do it?
- What is the next step?
- What do I still not understand?

In the Classroom

- Independent projects (peer assessment!)
- Reflect on assignments (quickly or in detail)
• Madison’s daily blog
• Use Google Doc to reflect (automatic updates back to you!)

questions
1) Are the high count genes in our DE 0% vs 190% regulon?
2) The correlations between the replicates of each sample are pretty good...even the spearman (.75 - .95). What do we think about that and what further analysis could I do? (Maybe I could look at distances by gene specifically, so I'm not just looking an aggregate figure).
3) Could large counts really be affecting size factors? I'm confused about that because I thought the median method helped disregard outliers in the creation of size factors (see blog post 6/23). The question: are size factors really systematically being pulled around by the inclusion of IGR, or not, for example?
4) How could analyzing dispersions help our analysis?

todo
1) Look at replicate distances across Geneid
2) How could large counts be affecting size factors? Or including different gene regions be affecting size factors?
3) Look up information about dispersion and try to understand this better. Also read R with Convincing!
4) Work through clustering analysis resources Dr. Hardin sent me!
5) Revise technical paper for Jonathan with feedback!
Wrangle Data

In the Classroom

- Data Science
- Statistics
- Theoretical

Practice, practice, practice

- Learn how to problem solve independently.
- Data wrangling should happen in every class at every level.
Data Wrangling

- dplyr
- ggplot
- grep

DataCamp
- Madison found a potentially important intergenic region
What else?

- Become a part of the larger community (e.g., quo() function in dplyr v 0.7.1, June 22, 2017)
- Use Git & GitHub (http://happygitwithr.com/)
- Bring your love of research to the classroom to generate excitement.
Thank you!

Jo Hardin
Pomona College
jo.hardin@pomona.edu
@jo_hardin47
Github: hardin47