Incorporating Data Analysis and Writing in Introductory Statistics

Rebecca Nugent, Philipp Burckhardt, Gordon Weinberg

Department of Statistics & Data Science
Carnegie Mellon University

AMATYC 2021
The Science of Data Science

A process or workflow; solving real problems by “extracting value from data”

Management includes security, elements of data engineering

Interpretation includes communication

In practice, move roughly from left to right but with loops and iterations; experts often focus on specific pieces; project managers oversee pipeline

J. Wing (2019), Harvard Data Science Review
The Science of Data Science

The Ultimate Choose Your Own Adventure Book (where hopefully the data analysis doesn’t lead to being trapped in a cave forever):

With apologies to Edward Packard
Open-Ended Data Analysis

Issues/questions we commonly hear:

▶ How do we know this is right?
▶ My friend’s graph looks like this; why does mine look different?
▶ My friend’s p-value is this; why is mine different?
▶ How do I decide what to do?
▶ Seriously, is this right?

(Intro) students don’t have enough exposure to the entire data science/statistical analysis pipeline.

Have little to no intuition for variation introduced by decision-making; how do we reconcile differences across analyses?
Carnegie Mellon University

- Private university in Pittsburgh, PA; R1 research university designation
- ≈ 7000 undergrads, 7000 grads
- Seven colleges: College of Fine Arts, Dietrich College of Humanities & Social Sciences, College of Engineering, Heinz College of Information Systems and Public Policy, Mellon College of Science, School of Computer Science, Tepper School of Business
- ≈ 550 primary/additional majors; Statistics (Concentration: Open, Math, Neuroscience); Economics-Statistics, Statistics and Machine Learning
- Almost all of our course sizes (UG through PhD) are in the hundreds
Where we were/Where we’re going

Interviews w/ different depts (and industries) about issues/needs

▶ Students don’t know the concepts
▶ Get tied to the specific software syntax or steps (Minitab)
▶ Can’t see the big picture
▶ Classes aren’t really for them

Our goals:

▶ Modernize courses; support different styles of learning (incl. remote)
▶ Emphasize concepts; tell stories with data
▶ More student-driven inquiry, case studies
▶ More adaptive material
▶ learn how different students interact with data

rnugent@stat.cmu.edu http://www.stat.cmu.edu/isle isle@stat.cmu.edu
Integrated Statistics Learning Environment (ISLE)

- Labs; Surveys; Widgets
- Sketch Pads, Interactive Lecture Slides; Group Collaboration
- Data Explorer; Reports; Presentations
- Peer to Peer Sharing; Chat Rooms
- Data Provenance; Reproducibility
- Searchable Action Logs; Grading/Feedback

rnugent@stat.cmu.edu  http://www.stat.cmu.edu/isle  isle@stat.cmu.edu
Integrated Statistics Learning Environment (ISLE)

- Browser-based; multiple operating systems and devices; can be downloaded for offline access
- Conformance to most of the World Wide Web Consortium’s Web Content Accessibility Guidelines (W3C WCAG 2.4)
- Active work on improving usability when using screen readers
- Integrated video & audio chatting through Jitsi meet

Enabling Technology: Capstone Collaboration in the ISLE Sandbox

- Interact with data / data exploration
- Share comments on data sets
- Communicate via chat with simple file sharing (images/pdfs)
- Collaborate on group reports with templates
- Work via shared notes/sketchpad functionality
- Communicate via real-time video
Integrated Statistics Learning Environment (ISLE)

Where is it?

- Freshmen through graduate students at Carnegie Mellon
- Community colleges, liberal arts, public univ, research univ
- Statistics/Data Science through English/Humanities classes
- Flipped classroom, remote learning, choose your own adventure
- Retraining/upskilling/ExecEd: health care, finance, manufacturing, etc
- Interactive journal article content

Research:

- Writing and structuring arguments about data
- How to optimize a data science team; group collaboration
- Populations and variance of data analyses (*Many Students, One Dataset*)
- Data literacy; longitudinal impact related to access and equity

rnugent@stat.cmu.edu  http://www.stat.cmu.edu/isle  isle@stat.cmu.edu
Integrated Statistics Learning Environment (ISLE)

https://isle.stat.cmu.edu/amatyc2021/sandbox
Data Science, Writing

- Currently used to support both individual and group writing
- Open-ended data analysis reports; exercises structured for style, language
- CMU Introductory Statistics and Writing About Data (English), general education courses for hundreds of students

Introduction to COCA

Using the Corpus of Contemporary American English (COCA)

The Corpus of Contemporary American English

Once you have registered for the Corpus of Contemporary American English (COCA) and joined the Carnegie Mellon University group, using the corpus is fairly intuitive. (You can also refer to the guide on Canvas.) However, it does have some specific features and capabilities. The purpose of this exercise is to get you acquainted with some of these, and also to help you to be thinking about patterns in language as a springboard for your project.

Basic searching

Here is the basic interface:

To search, you simply type a word or phrase into the field and hit “Find matching strings” (a string is a computer science-y term for any series of characters including spaces and punctuation). Also note that at the top, you have 5 options that will change the output of your search. For this exercise, we will be using the “List”, “Chart”, and “Collocate” views.

Part-of-speech tags

The underlines followed by letters are part-of-speech. So the search increase, *n* returns all uses of increase used as a noun (with the increase in volume), while increase, *v* returns all uses of increase used as a verb (the volume will increase). Alternatively, the search increase, *p* (with a space between the word and the tag) returns all prepositions that follow increase (increase in, increase of, etc.).

Also note that the square brackets around a word will search for all forms of that word. So [increased] returns all forms of increase regardless of part-of-speech (increase, increases, increasing, increased). We could combine tags and brackets as in [increased],*n* to find, for example, only forms of increase used as a noun.

Frequency charts

Now let’s try switching to the “Chart” view. If we were to type thing in the search field with chart option selected, we would get something like this:

http://www.stat.cmu.edu/isle

rnugent@stat.cmu.edu

isle@stat.cmu.edu
Data Science, a Team Sport

We want to:

▶ teach our students/practitioners how to do (good) data analysis
▶ teach them how to work together

Now we have different sets of decisions from different people with different backgrounds and somehow they’re supposed to agree on one final product.

So what happens? How do we do it?

▶ Have them sit and work in groups during lab/class
  *Some of them talk to each other; others stare at their laptop/paper; others wait until class is over*

▶ Have them work together outside of class
  *Completely impossible to find time that all 4-5 students can meet*
  *Students end up just dividing up work; no cohesion*
  *1-2 people do all the work and the others do the intro/conclusion*

▶ Have them use "shareable" tools: Google, Slack, etc
  *Easy to ignore groupmates but sort of works. Assumes that students are okay with downloading and running tools. Can get by without speaking to each other.*
Data Science, a Team Sport

The truth is: we don’t know what they’re doing. And we’re not doing a very good job of teaching them how to optimize working together.

We want to help them learn about data by using data.

https://isle.stat.cmu.edu/amatyc2021/sandbox

In ISLE:

▶ can be assigned to groups
▶ can chat with each other and the instructor/TA
▶ can view each other over video

On the back end:

▶ track all the actions, time stamps, you name it
▶ can analyze when and how people worked
▶ can tie to performance but also tailor instruction

rnugent@stat.cmu.edu  http://www.stat.cmu.edu/isle  isle@stat.cmu.edu
Data Science, a Team Sport
Data Science, a Team Sport

The Instructor View:
Data Science, a Team Sport

Challenges:

- How can we reduce the mental tax of having to switch between different contexts (chat, report writing, data analysis)?
- How can one avoid having some students not participating, leaving their groupmates alone in finishing their reports?
- When work independently on different sections, how can one synthesize the findings?

Current Work:

- Improved collaborative editing experience including a searchable history of analysis actions and text edits
- Full group projects which may be completed asynchronously over a period of time
- Giving the students feedback as they work
- Compare different kinds of group assignments

rnugent@stat.cmu.edu    http://www.stat.cmu.edu/isle    isle@stat.cmu.edu
Final Takeaways

▶ Building accessible tools that allow for more student inquiry and open-ended analysis
▶ Can help facilitate student writing while also giving feedback on their data analysis steps
▶ Helps facilitate students working in groups, remote or in-person
▶ People from different backgrounds might actually just be thinking about data differently (not incorrectly)
▶ Students love the engagement with data, particularly if they can collect/analyze their own interests

To learn more about ISLE, please feel free to contact us at isle@stat.cmu.edu

A reminder that all Saturday AMATYC sessions can be evaluated on the WHOVA conference app until 7pm.
So what are we learning/researching?

- IRB allows access to action logs, etc after the semester is complete. Students can opt-out (so far they’re mostly not).
- Everything tracked. Everything.
- **Writing and structuring arguments about data**
- How to optimize a data science team; group collaboration
- Populations and variance of data analyses ("Many Students, One Dataset")
- Data literacy; longitudinal impact related to access and equity
- Optimizing lesson/lecture structure wrt student engagement (*early*)
- Adaptive review repositories with random question generation (*early*)
- Examples from Fall 2017 Intro Stat \((n = 71)\); Spring 2018 \((n = 130)\) tens of thousands of actions, 11-12 labs, data analysis reports
Creating/Describing Graphs

Comparison word clouds via answer TF-IDF values (graph type; over time)
Open-ended Scenarios

Lab session in week five of class uses a single dataset about school absences in Portugal, consists of four question scenarios:

- **Scenario 1**: Number of absences by location, urban or rural?
- **Scenario 2**: Older students more likely to miss school?
- **Scenario 3**: Academic performance by number of classes failed, differences between males and females?
- **Scenario 4**: Relationship between age and alcohol use?

**Scenarios 1-3**: critique and write description with explicit instructions on what stats and graphs to edit/create

**Scenario 4**: only write description with no guidance

Refer to as: S1 Critique, S1 Description,..., S4 Description
Open-ended Scenarios

Design graph for research question, critique current answer, rewrite

Scenario #4: It is thought that there is a relationship between the age of the student and the level of weekday alcohol use. Specifically, the older a student, the higher the level of weekday alcohol consumption.

Your Description
Your answer:
Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.
Open-ended Scenarios

Design graph for research question, critique current answer, rewrite

Time: 11:30:22 PM | User: ryrko@andrew.cmu.edu
ID: description_scenario4 | Type: FREE_TEXT_QUESTION_SUBMIT_ANSWER
Value: Based on a scatterplot of weekday alcohol use against age, it appears to decrease as age increases except for 22 years old.

Time: 11:24:33 PM | User: ryrko@andrew.cmu.edu
ID: schoolabsence | Type: DATA_EXPLORER:SCATTERPLOT
Value: {
    "xval": "Age",
    "yval": "WkdyAlc",
    "color": null,
    "type": null,
    "regressionLine": false,
    "regressionMethod": "linear",
    "lineBy": null
}
Open-ended Scenarios
Cluster students by their TF-IDF values with spherical k-means
Incorporating Timelines

Understanding and Analyzing Analysis Pathways

Sunburst chart of any three subsequent action types in students' data analyses. The light-red shaded pieces correspond to histograms, purple for summary statistics, and light-green ones for two-sample z-tests.
Incorporating Timelines

Analyzing (and clustering) how people build and write data analysis reports
Incorporating Timelines

Analyzing (and clustering) how people build and write data analysis reports

- Editor actions
- Plot actions