Statistics 133: Concepts in computing with data

Instructor: Jarrod Millman
GSI: Karl Kumbier

Monday, June 9, 2014

Background

- Who has used the command line? Bash?
- Who has used R? Python?
- Who has experience in programming?
- Who has used a version control system? Git?

Websites

Go to course website regularly:

`http://www.jarrodmillman.com/stat133-summer2014`

Discussion site:

`http://piazza.com/berkeley/summer2014/statistics133`

Academic integrity

Code of conduct:
`http://sa.berkeley.edu/code-of-conduct`

- Free to discuss course matters with instructor, GSI, and fellow students
- Keep the code you write to yourself
- Make significant contributions to your group's work
- Questions: If you are uncertain as to whether something may be a violation of the code, ask the instructor

Academic integrity

Writing code is like writing a paper — your code should be your original work.

A violation will result in at least one of the following:

- 0 for the assignment
- F for the course
- Report to the Office of Student Conduct

Computing

"Computing's core challenge is how not to make a mess of it."
– Edsger W. Dijkstra, EWD1243a

Radical novelty 1

"The first radical novelty is a direct consequence of the raw power of today's computing equipment. We all know how we cope with something big and complex; divide and rule, i.e. we view the whole as a compositum of parts and deal with the parts separately. And if a part is too big, we repeat the procedure. The town is made up from neighbourhoods, which are structured by streets, which contain buildings, which are made from walls and floors, that are built from bricks, etc. eventually down to the elementary particles."
– Edsger W. Dijkstra, EWD 1036

Radical novelty 2

"The second radical novelty is that the automatic computer is our first large-scale digital device. We had a few with a noticeable discrete component: I just mentioned the cash register and can add the typewriter with its individual keys: with a single stroke you can type either a Q or a W but, though their keys are next to each other, not a mixture of those two letters. But such mechanisms are the exception, and the vast majority of our mechanisms are viewed as analogue devices whose behaviour is over a large range a continuous function of all parameters involved: if we press the point of the pencil a little bit harder, we get a slightly thicker line, if the violinist slightly misplaces his finger, he plays slightly out of tune."
– Edsger W. Dijkstra, EWD 1036

Practices

- Version control with Git/Github
- Unit tests
- Automation
- Look at your data!
- Think

Statistical thinking and the data analysis cycle

- Acquisition
- Cleaning
- Organizing
- Analyzing
- Simulating
- Reporting

REPL

Read-eval-print loop

- UNIX/Bash
- R
- Python

Filesystems

- rooted tree
- `pwd`, `ls`, `cd`