using R, RStudio, and Docker for introductory statistics teaching


mine@stat.duke.edu
@minebocek
mine-cetinkaya-rundel

mine çetinkaya-rundel
duke university
playing nice in the classroom


mine çetinkaya-rundel
mine@stat.duke.edu
@minebocek
mine-cetinkaya-rundel

duke university
first course in stats for non-majors (STA 101)

not calculus based

mostly social science majors

possibly only quantitative course these students take in undergrad

weekly lab session + in class activities using R
why R?
why R?

- free & open source
- powerful & flexible
- relevant beyond intro stat

why not R?

- challenge of teaching programming in addition to stats concepts
- command line more intimidating than GUI

unlike other software designed specifically for courses at this level
challenge of teaching programming in addition to stats concepts

don’t do any hands on data analysis
don’t do any hands on data analysis
disservice to everyone involved
challenge of teaching programming in addition to stats concepts
use a drag-and-drop type tool
III. Adding Proportions to Summary Table

For categorical variables, you should see the counts of each possible outcome of that variable in the **Summary Table**. To see the breakdown of proportions or percentages, follow these steps:

- Click on the **Summary Table** to highlight it, click on the “Summary” drop-down menu and select “Add Formula”. In general, whenever you click and select a **Fathom** object (such as a **Table**, **Graph**, or **Summary**) the menu at the top of the screen will change to give you options for working on that object.

- In the formula editor that pops up, type “rowproportion” (without the quotes) to see the row proportions or “columnproportion” to see the column proportions. Be sure to spell the names of the formulas correctly or else **Fathom** will give you an error. (If you spell the names correctly, they should change to a purplish color in your editor.)

- You will see that each cell in the **Summary Table** now includes numbers for multiple statistics. To see which numbers correspond with which statistics, simply look at the bottom of your summary table to see the order of the statistics or formulas within each cell.

- To delete (or change) a particular statistic from the table, you can double click on its name at the bottom of the **Summary Table**. In the formula editor, press delete (or make your changes) and then click “OK”.

---

IV. Creating Graphs of Categorical Variables

It is always useful and important to visualize the data you are trying to analyze. To do this in **Fathom**, follow these steps:

- Click on the **Graph** button in the **Toolbar** and drag and drop a blank **Graph** object onto the **Document**.
don’t do any hands on data analysis
disservice to everyone involved
challenge of teaching programming in addition to stats concepts
use a drag-and-drop type tool
there’s still a learning curve
command line more intimidating than GUI
command line more intimidating than GUI
how R?

- technical
- pedagogical
**getting started:**
“like a knife through butter”

- avoid local installation
- preinstalled & preloaded packages
implementation: phase 1

external (RStudio) solution

RStudio beta server

Gmail authentication a pain

Control over version / packages limited

keep the experience

university login

full control
implementation: phase 2
in-house solution

option 1: monolithic RStudio server instance
  scaling issues
  load prediction
  security consideration (large # of non-dept students)

option 2: personal VMs
  resource intensive
  duplication

option 3: docker lightweight (with many virtues of individual VMs)
  sandbox individual students
  spin up new servers on the fly as needed
user

duke login

map to their assigned custom Docker containers redirect to container host / port

RStudio-Host-1
port 49100 + homedir100
port 49101 + homedir101
port 49102 + homedir102 etc...

RStudio-Host-2
port 49200 + homedir200
port 49201 + homedir201
port 49202 + homedir202 etc...

RStudio-Host-3
port 49300 + homedir300
port 49301 + homedir301
port 49302 + homedir302 etc...

rsync homedirs

backup

server: 8 GB RAM, 2 CPU, 300 GB disk, 110 containers/server
reproducible: literate programming

toolkit

train new researchers whose only workflow is a reproducible one

don’t touch the raw data

keep track of all analysis steps

avoid copy-paste

= Literate programming in...
support: lots to less

start with templates including code and answers

slowly remove handholding
R Markdown learning outcomes (beyond reproducibility)

learn R

avoid the messy / frustrating console

built-in and consistent syntax highlighting

n <- 1000
p <- seq(0, 1, 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
R Markdown learning outcomes (beyond reproducibility)

- learn R
- avoid the messy/frustrating console
- built-in and consistent syntax highlighting
- code and output always together

```r
sim_streak <- calc_streak(sim_basket)
barplot(table(sim_streak))

median(sim_streak)
## [1] 0

IQR(sim_streak)
## [1] 1
```
R Markdown learning outcomes (beyond reproducibility)

- learn R
- avoid the messy / frustrating console
- built-in and consistent syntax highlighting
- code and output always together
- feedback + grading
- ambiguity removed

```r
sim_streak <- calc_streak(sim_basket)
barplot(table(sim_streak))

median(sim_streak)
## [1] 0

IQR(sim_streak)
## [1] 1
```
<table>
<thead>
<tr>
<th><strong>R Markdown learning outcomes (beyond reproducibility)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>learn R</strong></td>
</tr>
<tr>
<td><strong>feedback + grading</strong></td>
</tr>
<tr>
<td><strong>collaboration</strong></td>
</tr>
<tr>
<td><strong>avoid the messy / frustrating console</strong></td>
</tr>
<tr>
<td><strong>ambiguity removed</strong></td>
</tr>
<tr>
<td><strong>just share the Rmd</strong></td>
</tr>
<tr>
<td><strong>built-in and consistent syntax highlighting</strong></td>
</tr>
<tr>
<td><strong>code and output always together</strong></td>
</tr>
</tbody>
</table>
resources
designed to be adopted / adapted

OpenIntro
openintro.org

specific to my course

stat.duke.edu/~mc301

mine-cetinkaya-rundel
acknowledgements

mark mccahill, duke OIT
thank you!

comments / questions?

mine@stat.duke.edu
@minebocek
mine-cetinkaya-rundel