BIOF 339: Practical R

Instructor

Abhijit Dasgupta
Contact: via Slack or email (adasgupta+biof339@araastat.com)

Course description

The goal of this course is to introduce R as an analysis platform and tool for data science rather than a programming language. Throughout the course, emphasis will be placed on example-driven learning. Topics to be covered include: installation of R and R packages; command line R; R data types; loading data in R; manipulating data; exploring data through visualization; statistical tests; correcting for multiple comparisons; building models; generating publication-quality graphics; creating reports using RMarkdown. No prior programming experience is required.

Learning Objectives

- Run R and RStudio, making use of inherent R features
- Find and make use of the extensive packages (R add-ons) available for analyzing biological and other forms of data
- Load, manipulate, and combine data to make it amenable to further analyses
- Visualize data with extensive graphics capabilities of R (including ggplot)
- Use R to run statistical models and hypothesis tests and report results conforming to standards expected in scientific journals
- Write reports using the powerful `rmarkdown` package and its derivatives

Required computers and software

You are required to bring to each class a personal laptop running Windows, Mac OS X or Linux. You are also required to install the software R and the integrated development environment RStudio. Instructions for installing these are available on the Resources page (see links above).

Outline of the class

Date	Topic
September 11, 2019	Introduction to R, RStudio and RMarkdown
September 18, 2019	Data Structures in R (classes 5:30-7, 7-8:30)
September 25, 2019	R packages, data import/export, munging
October 02, 2019	Towards analytic data: Data Munging, continued
October 09, 2019	Data exploration through visualization
October 16, 2019	More data visualization and RMarkdown
October 23, 2019	Statistical analyses: Table 1, estimation and confidence intervals, and more ggplot
October 30, 2019	Statistical analyses: Classical hypothesis testing and computational inference

Date	Topic
November 06, 2019	Statistical learning: Regression models
November 13, 2019	More data munging with `purrr`: grouping, mapping and functional programming
November 20, 2019	Basic bioinformatics: Bioconductor and friends
November 27, 2019	No class (Thanksgiving)
December 04, 2019	Statistical learning: Cluster analysis and pattern recognition
December 11, 2019	Project presentations

Books and learning materials

There are no required books for this class. However, we will extensively refer to a few books available freely online and will serve as reading material and ongoing reference material for this course.

1. *R for Data Science* [R4DS] by Hadley Wickham and Garrett Grolemund (available online)

Communication

This class will communicate primarily via Slack. Please join the BIOF339 Slack channel using this link.

You will see two channels named *wed5-7_2019* and *wed7-9_2019*. Please join the channel corresponding to your section. I will be using Slack for broadcasting messages, answering questions and the like. If you have a question, you can directly message me on Slack if you like. Expect an answer within 24 hours.

Grades

Grades will be based on the following requirements:

1. Homeworks, available Friday after class, due by 11:59PM the following Tuesday. (50%)
    - No late homeworks, since solutions will be available Wednesday mornings
    - I'll score the top 10 homeworks for grade
2. Final project: A RMarkdown report/presentation demonstrating an end-to-end data analysis in R using your own data, from data ingestion to munging to analyses and graphics, with a brief introduction and conclusion (20%)
3. Class participation (20%)
4. Completion and submission of class exercises (10%, marked for completion)
    - These will need to be in basic RMarkdown, showing the problem and the solution. You can add a section for questions here that I can address in the following class or online. These will have to be submitted before you leave the classroom

Academic policy regarding plagiarism

The FAES Graduate School at NIH prides itself on providing quality educational experiences and upholds the highest level of honesty, integrity, and mutual respect. It is our policy that cheating, fabrication or plagiarism by students is
not acceptable in any form. If a student is found to be in violation of any, or all of the below, his/her credits will be forfeited, and he/she will not be allowed to enroll in future courses or education programs administered by FAES.

• Cheating is defined as an attempt to give or obtain inappropriate/ unauthorized assistance during any academic exercise, such as during examination, homework assignment, class presentation.

- Fabrication is defined as the falsification of data, information or citations in any academic materials.
- Plagiarism is defined as using the ideas, methods, or written words of another, without proper acknowledgment and with the intention that they be taken as the work of the deceiver. These include, but are not limited to, the use of published articles, paraphrasing, copying someone else's homework and turning it in as one's own and failing to reference footnotes. Procuring information from online sources without proper attribution also constitutes plagiarism.

See this link for FAES policy on academic plagiarism.