Group Study: Modern Statistics for Modern Biology

Spring 2020

Instructors:	Fosdick, Anderson, and Lyons	**Time:**	Th 3:00–5:00 PM
Website:	kind-neumann-789611.netlify.com	**Place:**	Weber 223H

Course Description: To fully understand and interpret the increasingly complex and sophisticated laboratory data of contemporary life sciences requires quantitative skills beyond the traditional statistical analyses of biological research based on, for example, descriptive statistics and hypothesis testing. This course aims to extend the conventional biological data analysis skill set to include computational methods and tools that provide a for modeling, simulation, visualization, and the integration of complex data sets such as those needed to quantify gene and protein expression profiles from -omics type measurements, the numbers and type of cells and the chemical composition of a biological sample from flow-cytometry, and to categorize and quantify the geometric features in high resolution histopathology image data. The R programming language and environment will be used to explore basic concepts in probability and statistics, clustering, networks and trees, image data, high-throughput count data, and supervised learning. Additionally, reproducible research and collaborative science tools will be employed with git version control and R Markdown-based tools.

Prerequisites: This course assumes some experience programming in R and some background in statistics (e.g., STAT 511). You can still be successful without one or the other, but in that case you will need to be willing to commit to putting in a bit more work than students with that experience.

References:

The main reference for this course is:

- Susan Holmes and Wolfgang Huber, *Modern Statistics for Modern Biology*, Cambridge University Press, 2019. https://www.huber.embl.de/msmb/#book-supplements

Other references that might be useful are:

- Garrett Grolemund and Hadley Wickham, *R for Data Science*, O'Reilly, 2017. https://r4ds.had.co.nz/
- Yihui Xie, Amber Thomas, and Alison Presmanes Hill, *blogdown: Creating Websites with R Markdown*, CRC Press, 2018. https://bookdown.org/yihui/blogdown/
- Rob Phillips, Jane Kondev, Julie Theriot, Hernam Garcia, *Physical Biology of the Cell*, 2nd Edition, Garland Science, 2012.
- Phillip Nelson, *Biological Physics: with New Art by David Goodsell*, Macmillan Higher Education, 2013.
- Phillip Nelson, *Physical Models of Living Systems*, W. H. Freeman and Company, 2015.
- Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Toberts, Peter Walter, *Molecular Biology of the Cell*, Garland Science, 2017.

Tentative Course Schedule:

Chapter 1: Generative Models for Discrete Data ..... Feb. 6
- Vocabulary: Faculty
- Exercise: Faculty (Exercise 1.8, p. 17)

Chapter 2.1-2.7: Statistical Modeling (part 1) ..... Feb. 13
- Vocabulary: Faculty
- Exercise: Sere Williams (Exercise 2.3, p. 50)

Chapter 2.8-2.12: Statistical Modeling (part 2) ..... Feb. 20
- Vocabulary: Sierra Pugh
- Exercise: Sarah Cooper (Exercise 2.6, p. 51)

Chapter 4: Mixture Models ..... Feb. 27
- Vocabulary: Amy Fox
- Exercise: James DeLisio

Chapter 5: Clustering ..... March 5
- Vocabulary: Burton Karger
- Exercise: Camron Pearce

Chapter 6: Testing ..... March 12
- Vocabulary: Daniel Dean
- Exercise: Sherry WeMott

Chapter 7: Multivariate Analysis ..... March 26
- Vocabulary: Zach Laubach
- Exercise: Sierra Pugh

Chapter 8: High-Throughput Count Data ..... April 2
- Vocabulary: Mikaela Elder
- Exercise: Burton Karger

Chapter 9: Multivariate Methods for Heterogeneous Data ..... April 9
- Vocabulary: Sere Williams
- Exercise: Daniel Dean

Chapter 10: Networks and Trees ..... April 16
- Vocabulary: Sarah Cooper
- Exercise: Zach Laubach

Chapter 11: Image Data ..... April 23
- Vocabulary: Camron Pearce
- Exercise: Mikaela Elder

Chapter 12: Supervised Learning ..... April 30
- Vocabulary: Sherry WeMott
- Exercise: Amy Fox

Chapter 13: Design of High Throughput Exp. and their Analyses . May 7
- Vocabulary: James DeLisio
- Exercise: To be determined

Class Structure and Expectations:

- **Homework/preparation:** Each week we will focus on one chapter of the book. It is expected that before coming to class, students will read the chapter, study the vocabulary words for the chapter, and implement/review the in-text R examples.
- **In-class schedule:**
  - Vocabulary quiz: Each class will start with a vocabulary quiz on a select number of the words from the chapter's vocabulary list.
  - Small group discussion: Students and faculty will be divided into small groups to discuss the chapter and think more deeply about the content. This is a time to bring up questions and relate the chapter concepts to other datasets and/or analysis methods you are familiar with.
  - Group exercises: In small groups, students will tackle exercises related to the chapter's content. Often these will be from the textbook exercises at the end of the chapter.
  - Wrap-up: We will reconvene as one group at the end to discuss topics that came up in small group discussion and things learned from the exercises.

Grading Policy: Weekly vocabulary quizzes (25%), Vocabulary blog posts (25%), Exercise solution posts (25%), Attendance and participation (25%).

Description of Graded Materials:

- **Vocabulary quizzes:** There will be a short vocabulary quiz at the beginning of each class based on the keywords from the chapter.
- **Vocabulary blog post:** Each student will be responsible for creating the vocabulary list once or twice over the course of the semester. This should be saved as a tsv (tab-separated values) file and sent to the faculty 9 days prior to when the chapter will be discussed in class (i.e. the Monday of the week prior to the chapter). The faculty will review the file, suggest changes and then give approval for the student to post the vocabulary. The vocabulary should be posted to the website by the Friday before the chapter is discussed in class.
- **Exercise blog post:** Each student will be responsible once or twice over the course of the semester for creating a blog post with the solution to the in-class exercises. Students may borrow material from other students, either based on our work together in class or my discussing with them outside of class, to create the solution. The solution should be posted before the next class meeting.
- **Attendance:** The value of a group study is in the collaborative learning environment. Thus, attendance is an essential part of participating in the class. We understand things come up, however it is expected you attend every class and come prepared.

Academic Honesty: Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. For more on Colorado State University's policies on Academic Integrity / Misconduct.