Introduction to Neural Data Science
Fall term, 2022
Monday / Wednesdays, 4:30-6 pm Hess 9-101 (+ hybrid / Zoom)
Mark Baxter mark.baxter@mssm.edu
Denise Croote denise.croote@mssm.edu

Course credit/grading: 3 credits, letter-graded.

60%: homework assignments and problem sets (we will drop the lowest grade)

20%: group project and presentation

20%: regular attendance and participation (if you miss many of the class sessions, your grade will suffer)

What are the goals of this course?

- Gain an understanding of concepts of probability, including classical probability and distributions of random variables
- Introduce approaches in data management for common types of neuroscience problems, including good version control
- Introduce and discuss concepts in experimental design, including randomization, blinding, and correlation vs causation
- Gain facility with null hypothesis statistical testing approaches for continuous and categorical data
- Introduce linear models for explanation and prediction
- Develop skills in statistical programming using R and other platforms common in neuroscience (Python, MATLAB)

Who is this course for?

This course is designed for first-year students in the Neuroscience PhD program who will be dealing with neurobiological data. Neuroscience research tends to use both observational and experimental approaches, and the kinds of data that are generated and analyzed are extremely wide-ranging, from behavioral experiments in mice to neuroimaging and genomics studies on large populations of humans.

Quantitative rigor is critical in neuroscience research for at least two reasons. First, technical advances in neuroscience have resulted in an explosion of the volume of data generated by many experimental approaches. Second, funding agencies including the National Institutes of Health have been concerned about "rigor and transparency" in biomedical research, resulting in policies about data sharing that require a strong foundation in statistical computing.

This course may also be useful for students in other programs (for example, biostatistics, genetics and genomics) that want an exposure to the types of problems encountered in designing and analyzing experiments in neuroscience.

What is the format of the course?

Each class will include presentations by instructors about particular topics as well as example analysis problems using a combination of simulated and real data sets. We will start with an introduction to R. Students that are already proficient with R can speak to the instructors for alternative assignments during this phase of the class. We will also introduce computing in Python and MATLAB, two other platforms that are commonly used in neuroscience research.

There is no textbook for the course, but we will make use of online resources including:

"Statistical Thinking for the 21st Century" by Russ Poldrack (https://statsthinking21.org/)

"R for Data Science (2nd edition)" by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund (https://r4ds.hadley.nz/index.html).

We will use a Slack group to facilitate discussion.

Data sets and example code will be hosted here:

https://jetsetbaxter.github.io/intro-neural-data-science-2022

<u>Course schedule and topics</u>

August 15	Introduction and course overview / installing R and R Studio / "why R?"
August 17	Basic R programming / dice and cards / probability
August 22	Functions in R / probability games
August 24	"Tidyverse" in R / importing and manipulating data
August 29	Summarizing and visualizing data using the tidyverse
August 31	Data management and cleaning, outliers
September 5	*Labor Day Holiday - no class*
September 7	R projects and becoming friendly with GitHub
September 12	Python and MATLAB
September 14	Random variables and distribution functions
September 19	Simulating data and exploratory data analysis
September 21	Experimental design, "rigor and reproducibility", correlation vs causation
September 26	Sampling, sample vs population
September 28	Introduction to statistical inference; type I and type II errors
October 3	Statistical inference on means (z-test, t-test, permutation testing)
October 5	Correlation coefficients - quantifying the relationship between variables
October 10	*Indigenous Peoples' Day Holiday - no class*
October 12	Introduction to data visualization (histograms, scatterplots)
October 17	Statistical inference on counts (Fisher's exact test, chi-squared tests)
October 19	Parametric vs. nonparametric statistical tests
October 24	Power analysis
October 26	Experimental design from start to finish
October 31	Regression analysis and linear models
November 2	Linear models with more than one predictor
November 7	Linear models with categorical predictors
November 9	Diagnosing problems with linear models and evaluating assumptions
November 14	*Society for Neuroscience meeting - no class*
November 16	*Society for Neuroscience meeting - no class*
November 21	Statistical inference on linear models
November 23	*Thanksgiving - no class*
November 28	Data curation and management / responsible, NIH-compliant data sharing
November 30	Tools for improving reproducible data analysis (Markdown, Quarto, Docker)

December 5	Special topics and review
December 7	Special topics and review
December 12	Presentations
December 14	Presentations

Homework evaluation

These assignments are designed to give you some brief hands-on practice with some of the approaches that we talk about in class. We will ask you to submit code and in some cases output / results. The primary grading criterion will be completion of the assignment and submission of code that works. You are welcome to discuss the assignments with other members of the class, and search engines are your friend for figuring out how to do things if you don't know, but everyone must submit their own completed assignment to receive a grade.

Final project and presentation

You may work in groups of 2 or 3 (self-selected within the class). We will provide each group with data from a hypothetical experiment. Your job as a group will be to plan and carry out an analysis of the data based on a specified question – for example, whether there are differences between genotypes in an electrophysiological variable – and then describe your analytic process including descriptive and inferential statistics as well as appropriate visualizations. You will present your analysis to the class as well as produce a writeup consisting of a description of the analysis approach as well as your results, in a format similar to what would appear in a published manuscript.

Helpful resources for R coding and statistics

"R For Data Science" by Hadley Wickham and Garrett Grolemund https://r4ds.had.co.nz

`Swirlr` a package that runs tutorials within R https://swirlstats.com/students.html

"learnR4free" https://www.learnr4free.com/en/index.html

STAT 545 https://stat545.com/ course materials developed by Jenny Bryan for an introductory data science class - covers aspects of R programming and Github

"Introduction to Modern Statistics" by Mine Çetinkaya-Rundel and Johanna Hardin https://openintro-ims.netlify.app/index.html https://openintrostat.github.io/ims-tutorials/

Coursera https://www.coursera.org/courses?query=r%20programming

edX https://www.edx.org/learn/r-programming