Neural Data Science
Spring 1 term, 2022
Monday / Wednesdays, 4:30-6:30 pm Hess 9-101 (except 2/7/22: Hess 5-101)
Mark Baxter & Erin Rich
mark.baxter@mssm.edu @markgbaxter
erin.rich@mssm.edu @erinLrich

Course prerequisites:
BSR 1707, 1708 (Neuro Core units 3 and 4)
Either: BIO 6400 (Intro to Advanced Biostatistics) OR BSR 1026 (Applied Biostatistics for Biomedical Research)
BIO 6300 (Intro to R) highly recommended

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

60%: homework assignments and problem sets (the best 7 grades from these assignments will contribute to your final grade)

20%: final 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?

- Discuss issues in experimental design and structure of statistical analysis that tend to occur in the context of behavioral and neurophysiological studies, with specific reference to issues that are likely to be encountered in preparing fellowship applications
- Get exposure to statistical analysis approaches that aren't part of an introductory biostatistics course, including time series analysis
- Increase fluency with good data analysis practices to generate reproducible data analyses and shareable code

Who is this course for?

This course is designed for students in the Neuroscience PhD program who have already begun to design experiments and collect data. Data analysis becomes more concrete when you are actually running experiments and dealing with the practicalities of real data. It is also helpful to have some ideas about statistical approaches and issues in experimental design before you've collected all your data, so that you are not doing a postmortem cleanup operation on things that you could have fixed before you started running the study.

We also will be talking about issues that tend to be encountered in preparing fellowship applications, including how to compose a power analysis that will pass muster in the context of NIH "rigor and transparency" criteria.

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

What is the format of the course?

We will tackle different problems each week using a combination of simulated and real data sets. We also encourage participants in the course to bring their own data, either final or from pilot experiments, or problems with experimental design and/or data analysis to the group for discussion.

We will use a Slack group to facilitate discussion.

Data sets and example code will be hosted on here:

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

What this course is not

We expect you will have some programming / coding background before the class begins. This class is not a substitute for a class in R programming.

You should be able to run R on your computer, read in data to R, and run (for example) a two-sample t-test on it. We will briefly review some approaches for data management but if you have never analyzed data in R before you will be at a disadvantage.

You should have a background in basic statistics, including fundamentals of probability, probability distributions (normal, t, chi-squared, F, binomial, Poisson distributions), statistical inference (type I and II error, confidence intervals) and statistical tests for one- and two-sample data (t-tests, chi-squared tests). **This course is not a substitute for introductory courses in programming or biostatistics.**

In the format of a short, discussion-style course, it is not possible to fully develop the theory behind statistical approaches such as linear mixed models. The intention is to introduce these models and go through some practical examples, so that if they are applicable in your own work they are demystified enough that you can work on implementing them and not fall back on reducing everything to a t-test. **If your work depends heavily on sophisticated statistical approaches, you would be well served by taking additional formal courses in biostatistics and/or involving a biostatistician in your research program.**

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

Dataquest.io http://dataquest.io/ (we will provide free account for all students)

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

"Statistical Thinking for the 21st Century" https://statsthinking21.org/ online, open source textbook for statistical analysis by Russ Poldrack. Includes R and Python programming code

"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

"Analyzing Neural Time Series Data" by Mike Cohen https://mitpress.mit.edu/books/analyzing-neural-time-series-data

What should I do to prepare for the course?

Everyone should have familiarity with R, for example from prior course experience (e.g., BIO 6300 or BIO 6400). We do not expect everyone to be expert R coders, but you will be at a serious disadvantage if you don't have basic facility with R.

Course schedule and topics

January 19	Introduction / refresher on probability and statistical inference **HW0 due (pretest)**
January 24	Quick intro and refresher on R
January 26	Data wrangling in R / "tidyverse" tools
January 31	Power analysis theory and application **HW1 due (data wrangling)**
Feburary 2	Linear models for analysis and prediction
Feburary 7	More linear models / multilevel models **HW2 due (power analysis) (5-101)**
Feburary 9	Generalized linear models, logistic regression
February 14	Model comparison and inference, hierarchical models **HW3 due (linear models)**
Feburary 16	Survival analysis / time-to-event data
February 21	*Presidents Day Holiday - no class*
February 23	Graphics and data visualization **HW4 due (MLMs)**
February 28	Crash course in MATLAB
March 2	Introduction to time series analysis **HW5 due (graphics and visualization)**
March 7	Understanding frequency decomposition
March 9	Functional connectivity **HW6 due (time series)**
March 14	Population coding
March 16	Data mining part 1 **HW7 due (frequency decomposition)**
March 21	Data mining part 2
March 23	RNA-seq data analysis (Li Shen) **HW8 due (population coding)**
March 28	*Spring break - no class*
March 30	*Spring break - no class*
April 4	Discussion / catchup / special topics **HW9 due (data mining)**
April 6	Presentations
April 11	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.

Specific topics to be covered

- **Introduction / refresher on probability and statistical inference:** probability distributions, interpretation of statistical tests (Type I and II error), nominal/ordinal/interval/ratio scales, common statistical tests **(**https://bookdown.org/roback/bookdown-BeyondMLR/ch-distthry.html**)**
- **Data wrangling in R and "tidyverse" tools:** Common issues with bringing data into R, tips for working with data in different formats, never edit the primary data, bring example data files / formats to class for discussion
- **Power analysis theory and application**: Tradeoff between type I error and power, what is most likely outcome of an experiment where you don't know what the outcome will be, "dance of the p-values", power analyses for standard statistical tests, multiple comparisons, simulation approaches, what do you need to put in your grant
- **LInear models for analysis and prediction**: basic linear regression model, extension to multiple regression, regression vs ANOVA; type I versus type III sums of squares
- **Multilevel models for nested data**: Importance of separating variance within unit of observation versus between units of observation (e.g. neurons nested within animals); why does this matter and what happens if you do not appropriately model variability
- **Multilevel models for repeated measures data**: Extension of multilevel model to longitudinal data
- **Generalized linear models, binomial and logistic regression**: Illustration of modeling outcomes that are binary or discrete (Poisson regression), why does this matter
- **Model comparison and inference, hierarchical models**: evaluate contribution of adding factor of interest over "nuisance" variables, refine predictive models
- **Graphics and data visualization:** ggplot tips and tricks, best practices, what makes a bad graph https://twitter.com/biogeobiochem/status/1172547846479831040
- **Introduction to time series analysis:** types of data (spiking, oscillations, Ca++ imaging, fMRI, fiber photometry, etc., also how concepts can extend to non-brain data such as video tracking), time alignment, smoothing, normalization, quantifying “selective responses”
- **Understanding frequency decomposition:** time domains and frequency domains, approaches to frequency decomposition, understanding phase, amplitude, and power, ERPs and phase resets
- **Functional connectivity:** spike and phase coherence, cross-frequency coupling, measures of directional connectivity, graphical models of functional connectivity
- **Population coding:** understanding and quantifying distributed coding, understanding and implementing decoding approaches, introduction to neural network models
- **Data mining (2 parts):** approaches to exploratory data analysis (dimensionality reduction, classification, clustering analyses, visualization), potential pitfalls and best practices