Week 4: Data analysis application

Commentary on this assignment

This document is available for comments and questions. Everyone in the class can comment on it, point out errors, ask clarifying questions, and add helpful resources. I will try to update things, fix mistakes, and clarify throughout the week.

The goal of this assignment is to help you practice the process of exploring a dataset, posing relevant questions, determining what tests/methods to use, analyzing data, visualizing the data, reporting your findings, and finding statistics resources.

Given the long time between now and the next class, I am setting an interim deadline for choosing a dataset and developing preliminary questions (the goal of this is to help you commit to a dataset early, though you can always add an additional dataset and to give you time to pivot if needed). This assignment has 2 deadlines (one for submitting/describing your data set and one for the actual analysis).

Note: This assignment has a minimum expected engagement time of 5 hours. This is definitely not a maximum or suggested engagement time. Since people are coming in with really different experience levels with data analysis (which is great), I expect that you'll get to different depths of analysis. Please challenge yourself to learn something new or really reinforce something you've encountered already.

We'll talk about machine learning and neural networks briefly in the next part of the course, so I generally suggest avoiding them for this assignment, but talk to me if you are bursting with enthusiasm to do some machine learning on this assignment.

Your mission, if you choose to accept it (aka, the assignment)

1) Choose a dataset (or potentially a series of datasets) about something you find interesting. [Due early, see Canvas]
    a) The dataset does not need to be neurotech related, it might even be better to tie this to another topic so you can practice generalizing statistics outside of the neurotech context, but I'm very flexible here.
    b) Try to challenge yourself by choosing something that goes beyond a t-test (unless you're still really struggling with t-tests). For example, you can start to look at what happens when you have more than 2 conditions (ANOVA), linear relationships (linear regression), comparing frequencies/counts (X-squared

tests), Bayesian updating to estimate probabilities of hypotheses. This may require additional resources beyond what we've covered in class.

   c) You can opt to choose a paper that provides a data set and replicate their analysis. If you use this strategy, you might learn a lot from testing the effects of making small tweaks to the analysis (e.g, excluding certain data points or choosing a different statistical threshold).

2) Explore the data, figure out questions to evaluate, analyze the data (these guidelines may be helpful).
3) Report your analysis in a clear, well-documented way (a code notebook or a short writeup with figures and numbers).

Dataset resources

- Open Psychology journal, where each article includes a link to the original dataset. You can search for articles in the top right or browse the latest and most popular articles
- Raw data from online personality tests
- Collaborative Research in Computational Neuroscience
- Data sets from a Quantitative Methods in Neuroscience course (these give suggested analysis methods (many of the datasets are fictitious)
- Google dataset search (woah, you can search google for datasets now)
- Dr. Rasp's Data Sets for Classroom Use (Various non-neuro topics)
- You can add more suggestions here by commenting on this line

Statistics resources

- Statistics in Matlab textbook
- Statistics in the Computer Age textbook
- MIT OpenCourseware 18.05 stats course
- Medium, toward data science, and other articles
- You can add more suggestions here by commenting on this line