Deeper Data Analysis for Neuroscience and Psychology

John P. McGann
Spring Semester, 2020
Rutgers University

Course Information:
Title: Psych 533 Experimental Design and Methods: Deeper Data Analysis for Neuroscience and Psychology
Rutgers Course Number: 16:830:533:01
Date and Time: Thursdays 10:00-12:30
Index Number: 33785
Location: Psychology Building Room 301
Credits: Three
Prerequisite: Introductory statistics experience

Instructor:

Dr. John P. McGann
Email: john.mcgann@rutgers.edu
Office: Psych 311 (Busch Campus)
Office Hours: Thursdays 12:30-1:30 (after class)

Introduction and Goals of this Course

In neuroscience and psychology, traditional training in statistics focuses on assessing experimental outcomes using parametric null hypothesis testing, including p-values, t-tests, and ANOVAs. However, powerful alternative methods exist for avoiding parametric assumptions, for comparing experimental data to (non-null) hypotheses, and for performing exploratory data analyses in multivariate datasets using statistical and machine learning approaches. These methods are becoming an increasingly important part of the scientific toolbox. The present course is meant to provide practical training in these alternative approaches along with a big picture perspective on when and why to employ them. Because this course is intended for scientists, we will spend rather less time on mathematics underlying these established techniques and instead make sure you understand them conceptually and know how to actually use them. The course will be organized around pairings of concept and implementation as detailed below.

Prerequisites

Statistical Training: Students are expected to have previously taken an introductory statistics course (undergraduate or graduate) and thus be familiar with the logic of correlation, regression, t-tests, and analysis of variance. These subjects will not be covered except as a contrast with alternative methods.

Programming Experience (optional): For this initial offering of the course, students are not required to have coding experience. However, by mid-semester we will be performing machine learning analyses that are beyond the scope of conventional statistics packages. Students will need basic skills using a scientific programming environment like Matlab or Python to execute these tasks, but will be using pre-

built sections of code. This should be readily achievable for novices, though students with more advanced skills will also benefit.

Software
The course will be taught in a software-agnostic way. Early parts of the course will employ Origin (a powerful alternative to SPSS that is available for free through a Rutgers site license at software.rutgers.edu) for non-parametric statistics, data display, and curve-fitting. Later parts of the course will use Matlab for multivariate statistical analysis and machine learning approaches, including the distribution of Matlab sample code. Students are welcome to instead use alternative software throughout the course (e.g. Python or R) if desired, with the understanding that this may limit the technical feedback they receive from the instructor.

Evaluation
Problem sets: To ensure hands-on skill-building, students will be assigned problem sets corresponding to each course module. Most include implementing specific analyses on a provided dataset, though some will include conceptual questions. These problem sets will account for 70% of the final grade. Some of these problem sets will be initiated during class time. In completing these problem sets students will sometimes be afforded the option to substitute their own datasets for the provided ones but do so at their own risk.

Conceptual exam: At the end of the semester, students will complete a written, non-computational exam on the key concepts of data analysis and statistics. This exam will count for 30% of the grade.

Course Content

Unit 1: Moving beyond introductory statistics (Weeks 1-6)

Concept: Null-hypothesis testing vs. hypothesis specification

Implementation: Formulating precise, scoped hypotheses

- Qualitative vs quantitative hypotheses
- Causal vs non-causal hypotheses
- Defining the scope of a hypothesis (e.g. subjects & contexts)

Concept: Data distributions as things-in-themselves vs clues to a noumenological truth

Implementation: Data characterization and visualization

- Univariate scatterplots vs bar/box plots
- Histograms, cumulative frequency plots, probability distributions
- Bivariate scatterplots
- Multivariate scatterplots: color, point size, point type
- Scaling (log, linear, etc.), normalizing, data slicing, and other choices
- Editorializing via data display

Concept: Non-parametric statistics: what and why

Implementation: Performing basic non-parametric testing

- Kolmogorov-Smirnov testing
- Mann-Whitney, Wilcoxon, and Kruskal Wallis tests

- Friedman's ANOVA

Concept: Fitting data to quantitative hypotheses

Implementation: Fitting curves to data

- Goodness of fit measurements
- Parameter estimation

Concept: Choosing between data fits

Implementation: Applying Bayes' rule to quantify relative likelihood

- Log-likelihood ratios
- Bayes Factors
- Priors and empirical Bayes

Concept: Fitting data to qualitative hypotheses

Implementation: Statistical classification

- Chi-square and its limits
- Naïve Bayes classifiers
- Support vector machines

Concept: Sample independence & replication

Implementation: Making reasoned decisions about non-independent measurements

- Bootstrapping and statistical inference
- Defining replication within an experiment, lab, and field
- Experimental controls as a mixed blessing

Unit 2: Exploration and Hypothesis Testing in Multivariate Datasets (weeks 7-13)

Concept: Multivariate representations of complex data

Implementation: Representation of data in multidimensional space

- Understanding distances and vectors in N-dimensional space
- Understanding clustering in N-dimensional space

Concept: Dimensionality reduction via hidden variables

Implementation: Principal components analysis and Independent components analysis

- Performing PCA/ICA
- Quantifying goodness of fit (i.e. variance captured)
- Displaying data in principal component space

Concept: Categorization of data via supervised algorithms

Implementation: Manual and machine learning approaches and validation

- Manual cluster cutting
- Naïve Bayes classifiers
- K-Nearest Neighbor
- Random forest regression

- Support vector machines

Concept: Categorization of data via unsupervised algorithms

Implementation: Unsupervised machine learning and validation

- Hierarchical clustering
- K-means clustering
- Self-organizing maps
- Gaussian mixture modeling

April 23: Final Class: Brief overview of where to go from here (week 14)

Time varying data

- Autocorrelation
- Reverse correlation
- Manifolds in N-dimensional space, metric tensors, and tensor fields
- Markov models
- Limit cycles & non-linear dynamics

Different data types

- Spike trains and other neural data
- Genomics
- Video analysis
- Audio analysis

More sophisticated machine learning approaches

- Matlab Deep Learning Toolbox
- Keras/Tensorflow/PyTorch

Information based analyses

- Mutual information and joint distributions
- Granger "causality"

Course Schedule (tentative)

January 23 – Introduction, Hypotheses
January 30 – Data Visualization
February 6 – Non-parametric statistics
February 13 – Fitting data to quantitative hypotheses & Choosing between fits
February 20 – Sample independence and replication
February 27 – (Kleinschmidt lecturing) Bayesian vs frequentist approaches to inference
March 5 – Multivariate analysis
March 12 – Dimensionality reduction
March 19 – **Spring Break – No Class**
March 26 – Categorization via supervised algorithms
April 2 – Categorization via supervised algorithm
April 9 - Categorization via unsupervised algorithms
April 16 - Categorization via unsupervised algorithms
April 23 – Final class, overview of other methods
April 30, 2020 - **Conceptual Exam**