Final Exam Review Lecture Notes

Extract Information from Data

Data can be from any source

Competing Paradigms

- Paradigm 1: have research question/hypothesis, experimental design, get data, test question, report
- Paradigm 2: have data already, explore to find insights/patterns (be careful for multiple testing)

Importance of understanding data generation mechanism

- In paradigm 1, not a big deal because we generate our own data
- Paradigm 2, you probably didn't generate data, so make sure you understand how it was generated

Statistical Concepts vs. Programming Concepts

Lots of time with R

- Basic structures / Data Types / Vectors, Lists, Data Frames, Matrices
- Subsetting
- Boolean Logic
- Creating and Manipulating Vectors (rep, seq)
- Reading in Data
- Apply Family

Linear Methods

- Supervised vs. Unsupervised Learning (Do you have labeled data?)
- Linear Regression
- Randomness in Data
- Probability Distributions (Normal, etc.)
- Confidence Interval
- Hypothesis Test (Null vs. Alternative)
- Outliers
- Correlation

Writing functions in R

- Inside or Outside of Scripts
- Documenting Functions / Exception Handling
- Anonymous Functions (Apply Family)
- Flow Control
    - For loops

    - If/Then
    - While

Data Lifecycle

- Reproducibility of findings
- Metadata

Simulation

- see slide 13 of “8_Simulation.pdf”

Graphics

- When/Where do graphics fit in data lifecycle
- R: Data types:
    - Quantitative
    - Qualitative
- Color Choices
    - Brewer’s Color Palette
- Best practices for plotting
- Improving Plots

Regular Expressions

- Literals/Classes/Modifiers

Web Scraping

- More tools for reading data into R
- JSON, XML, HTML, CSS
- XML
    - Trees
    - Xpath
    - Creating XML trees in R
    - KML

Workflow Tools & Data Science Lifecycle

Unix & Shell Scripting

- sed
- awk

Combine tools and methods in the data science lifecycle for reliable findings