Foundational Statistics // Bi 610

See the schedule for topics by week and links to slides.

Instructors:

Dr. Clay Small, csmall@uoregon.edu Dr. Bill Cresko, wcresko@uoregon.edu

Course Information

Virtual Office Hours: Tu/Th 12:00 PM – 1:20PM (Zoom)

Description of the course

This course in an introduction to data management, data visualization, and statistical inference. It is intended for early-stage graduate students with no background in statistics. No prior coursework (undergraduate or graduate) in statistics or programming is assumed. The primary objective of the course is to get students up to speed with respect to organization, manipulation, visualization, and analysis of data, using the R statistical language. The emphasis on application is strong, with the goal of enabling students (after the course) to analyze their own data sets with confidence using reasonable approaches, and, when faced with more difficult analysis problems, to be able to communicate their inference objectives clearly to expert analysts. Students will learn to organize and analyze data sets in the form of RStudio projects, using R Markdown files to reproducibly capture and render code, visualizations, and analyses. In-class exercises will be delivered in the form of pre-formatted R Notebooks, which can be interactively executed by students without having to write all code from scratch.

The course is designed to acquaint students primarily with univariate (single response variable) analysis. Multivariate analysis will be covered in the Advanced Biostatics 2-course series offered during the Fall and Winter terms. Examples and assignments in class will include data sets primarily from the biological sciences, including studies of morphological and molecular traits, behaviors, ecological questions, and clinical studies. For specific statistical topics covered in class, please see the course goals and tentative schedule below.

Course goals:

- Properly organize and format primary data and metadata files for analysis
- Learn programming fundamentals of the R statistical language, including objects, functions, iteration, and simulation.
- Make publication-quality data visualizations, including scatterplots, boxplots, frequency distributions, mosaic plots, etc.
- Understand Type I and Type II statistical error, including p-values and power analysis.
- Understand ordinary least-squares regression and linear models in general •
- Learn the fundamentals of strong experimental design
- Learn to apply general linear models to basic univariate analysis problems, including Analysis of Variance (ANOVA)
- Learn nonparametric approaches to parameter estimate and statistical inference, including resampling (bootstrapping), permutation, and rank- based analysis.
- Understand how to analyze binary response variables and frequency-based (e.g. contingency table) data sets.

Course Organization

Short exercises: During the course of the class, you will gain practice via small workbook-style assignments. It is anticipated that you will be able to complete most of these in one short sitting. While these assignments will not be graded per se, students will receive participation points based on their commitment to these exercises during class, through a weekly Canvas post (< 1 page), which may include a summary of what they learned from the exercises - 10% of grade

Problem sets: Students will be assigned five problem sets to complete independently. These will mostly focus on one or a few data sets each, and the goal will be for the students to organize, visualize, analyze, and interpret the data sets in light of specific scientific motivations. - 90% of grade

Preparing and submitting assignments: The information required to complete all in-class assignments and problem sets will be given in instructions on Canvas. Students should carefully follow the detailed instructions associated with each assignment. Students are encouraged to work together and share information. In particular, some students will have a higher skill level than others, and we encourage those students with more experience to help their peers. *However, no direct sharing of code is allowed – each student must write their own code.*

Assignments will be submitted on Canvas in the form of R Markdown files, with the rendered html output. Be sure to include your name on the document. Be professional – appropriately name your files, make sure they are organized, and submit only the information requested. Late assignments will not be accepted.

Textbooks

There is no formal textbook for this course, but supplementary reading will be suggested (depending on topic) for students seeking more information.

One reference you might find helpful is: * Logan, M. 2010. *Biostatistical Design and Analysis Using R.* Wiley-Blackwell. A fairly comprehensive book that covers how to use R

Software:

- Latest version of R (install here)
- Latest version of RStudio (install here)
- A terminal that allows ssh connection to the UO computing cluster (Talapas)

Prerequisites:

None. A practical knowledge of basic algebra may be helpful.

Inclusion and accessibility

Please tell us your preferred pronouns and/or name, especially if it differs from the class roster. We take seriously our responsibility to create inclusive learning environments. Please notify us if there are aspects of the instruction or design of this course that result in barriers to your participation! You are also encouraged to contact the Accessible Education Center in 164 Oregon Hall at 541-346-1155 or uoaec@uoregon.edu.

We are committed to making this course an inclusive and respectful learning space. Being respectful includes using preferred pronouns for your classmates. Your classmates come from a diverse set of backgrounds and experiences; please avoid assumptions or stereotypes, and aim for inclusivity. Let us know if there are classroom dynamics that impede your (or someone else's) full engagement.

Because of the COVID-19 pandemic, this course is being delivered entirely remotely. We realized that this situation makes it difficult for some students to interact with the material, for a variety of reasons. We are committed to flexibility during this stressful time and emphasize that we will work with students to overcome difficult barriers as they arise.

Please see this page for more information on campus resources, academic integrity, discrimination, and harassment (and reporting of it).

Foundational Statistics // UO Bi 610

Weeks 1-2

1. Data organization and management
    - best practices, reproducibility, etc.
2. Basic programming fundamentals for data curation
    - The Unix environment and fundamental commands
    - Formatting and manipulating tabular text files from the terminal
3. Introduction to R and Rstudio
    - Installation/Updates
    - R object types and assignment
4. Practice with R objects
    - vectors, matrices, data frames, etc.
5. Applying core programming fundamentals in R
    - vectorized operations
    - replicate, apply family, ifelse, for loops, etc.

Week 3

1. Plotting/visualizing data as a means of exploration
    - Different plot types
    - Scale, transformations, etc.
2. Fundamentals of plotting in base R
    - par
    - using palettes, points, sizes, etc. to convey information
    - axes and labels
3. R markdown

Week 4

1. Population parameters, samples, and sampling distributions
    - Central Limit Theorem and the normal dist.
    - Mean and st. dev.
2. Probability and probability distributions
3. Calculating summary statistics
    - Other common summary statistics (quantiles, etc.)

Week 5

1. Parameter estimation
    - Simulating data sets with known parameters
    - Revisit probability distributions
2. Uncertainty in estimation
    - Parametric and nonparametric approaches to uncertainty

Week 6

1. Experimental design
    - lexicon
    - considering sources of variance
    - types of variables (categorical, ordinal, rational)
    - confounding variables
2. Frequentist hypothesis testing
    - error types
    - p-values
    - degrees of freedom
    - statistical power
    - multiple testing problem

Week 7

1. Comparing means between groups
    - Student's t-test
2. Bootstrapping and randomization to compare means

Week 8

1. Relationships between quantitative variables
    - correlation and covariance
2. Simple linear regression
    - residuals and least squares
    - fitting linear regression models

Week 9

1. Analysis of variance
    - Table components and test statistics
2. General linear models in R
    - Model formulae
    - Interpretation of summary output
3. More complex ANOVA frameworks
    - Nested models
    - Factorial models

Week 10

1. Frequency-based statistical tests
    - Chi-squared tests
    - Contingency tables and tests of independence
2. Brief introduction to generalized linear models (time permitting)
    - logistic regression