Introduction to Biostatistics and R

Course Objectives:
The goal of this course is for students to develop a basic foundation in biostatisics and data analysis in observational and experimental studies. The ultimate intent is to provide students with statistical techniques and understanding that will permit them to learn more advanced statistical tools and application. At the end of this course, students will be expected to understand how to organize and clean their data and analyze and graphically present their results in R. The course is geared towards undergraduates in evolutionary biology and ecology and it will meet three times per week: twice for lectures and once in the computer lab to learn R.

In addition to being able to display and analyze results, students should have mastered concepts on experimental design, coding in R, choosing a model, and communicating results. Students will work on a semester long project and will be required to submit a paper and present results at the end of the course.

Lectures will be held on Mondays and Wednesdays from 11:00a-1:00pm each week. Computer lab sessions will be held on Thursdays from 1:00p-2:30p.

Resources:

- Whitlock, M.C & Schulter, D. *The Analysis of Biologicial Data, 2nd edition*. 2015. Roberts and Company Publishers, Inc., Greenwood Village, CO. (W&S)
- Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 2007. Cambridge University Press, New York, NY. (AH) (**There are many copies available at the library.**)
- *Optional:* James, G., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. 2013. Springer Science+Business Media, New York, NY. (JWHT)

Additional Reading: This course relies mainly on the textbooks but for some lectures, additional primary literature or online videos will be available for students to read or watch if they are interested. These optional reading and video assignments will be denoted by an asterisk (*).

Computer Lab Sessions: Each week, we will meet in the computer labs for an hour to learn how to use the statistical software R. The aim is to provide students with a firm foundation in the basic R packages, data cleaning, analysis and presenting their data. Students will need to first download R (https://cran.r-project.org/) and RStudio (https://www.rstudio.com/products/rstudio/download/) and then install the package Swirl. We will cover how to do these things in the first computer lab session.

SCHEDULE

WEEK 1: Introduction to Statistics

Meeting 1: What is statistics?

W&S: 1-17 (*Statistics and samples*)

*W&S: 23-24 (*Biology and the history of statistics*)
Meeting 2: Displaying and Describing Data
W&S: 26-52 (*Displaying data*)
W&S: 65-85 (*Describing data*)
Lab 1: Introduction to R
*JWHT: 42-51 (*2.3 Lab: Introduction to R*)
Swirl: R Programming - Basic Building Blocks & Workspace and Files

WEEK 2: Uncertainty and Probability
Meeting 3: Uncertainty and Hypotheses
W&S: 95-107 (*Estimating with uncertainty*)
Standard deviation vs. standard error
(https://www.r-bloggers.com/standard-deviation-vs-standard-error/)
Meeting 4: Probability
W&S: 117-138 (*Probability*)
*W&S: 115-116 (*Pseudoreplication*)
Lab 2: Building and Cleaning Data
Swirl: R Programming - 'Sequences of Numbers' and 'Vectors'.
Install the 'dplyr' and 'tidyr' packages before class.

WEEK 3: Hypotheses and Proportions
Meeting 5: Hypothesis Testing
W&S: 150-168 (*Hypothesis testing*)
Hypothesis Testing (P-Value approach) (https://onlinecourses.science.psu.edu/statprogram/node/138)
Meeting 6: Proportions and Frequencies
W&S: 179-193 (*Analyzing proportions*)
*W&S: 201-202 (*Correlation does not require causation*)
Lab 3: Working with missing values and subsetting
Swirl: R Programming - 'Missing Values' and 'Subsetting Vectors'

WEEK 4: Distributions and Inference
Meeting 7: Normal Distributions
W&S: 273-293 (*The normal distribution*)
*AH: 13-26 (*Concepts and methods from basic probability and statistics*)
Meeting 8: T-Tests and Inference
W&S: 303-318 (*Inference for a normal population*)
W&S: 301-302 (*Controls in medical studies*)
Lab 4: Matrices and Functions
Swirl: R Programming - 'Matrices and Data Frames' and 'Functions'

WEEK 5: Choosing a Test and Handling Assumptions
Meeting 9: Mean Comparison
W&S: 327-353 (*Comparing two means*)
W&S: 366-367 (*Which test should I use?*)
Meeting 10: Handling Assumptions

W&S: 369-400 (*Handling violations of assumptions*)
Type I and II Errors and Significance Levels
(https://www.ma.utexas.edu/users/mks/statmistakes/errortypes.html)

Lab 5: Using t-tests in R
YouTube: 'One-Sided Test or Two-Sided Test?' (https://www.youtube.com/watch?v=VP1bhopNP74)
Load the 'Iris' dataset in R before class (come early to class if you need help with this)

WEEK 6: Designing Experiments

Meeting 11: Sample Size and Effect Size
W&S: 423-450 (*Designing experiments*)
Lemoine, N.P., Hoffman, A., Felton, A.J., Baur, L., Chaves, F., Gray, J., Yu, Q., & Smith, M.D. (2016).
Underappreciated problems of low replication in ecological field studies. *Ecology, 97(10,)* 2554-2561.

Meeting 12: Experimental Design and Power Calculations
AH: 437-454 (*Sample size and power calculations*)
Noordzij, M., Tripepi, G., Dekker, F.W., Zoccali, C., Tanck, M.W., & Jager, K.J. (2010).
Sample size calculations: basic principles and common pitfalls. *Nephrology Dialysis Transplant, 25,*
1388-1393.

Lab 6: Module in Effect vs. Sample Size and Power Calculations
Download the script from the course website 'EffectvsSample.R'

WEEK 7: Understanding ANOVAs

Meeting 13: One-way ANOVAs
W&S: 459-471 (*The analysis of variance*) & (*Assumptions and alternatives*)
AH: 487-490 (*Analysis of variance: Classical analysis of variance*)

Meeting 14: Two-way ANOVAs
W&S: 471-486 (*Comparing means of more than two groups*) Part II
W&S: 500-501 (*Experimental and statistical mistakes*)

Lab 7: Intro to ANOVAs in R
YouTube: 'How to Calculate Anova Using R'
(https://www.youtube.com/watch?v=fT2No3Io72g)
Install the 'car' package - we will use the 'Anova' function rather than 'aov')

WEEK 8: Linear Regression Part I

Meeting 15: Intro to Linear Regression
W&S: 539-557 (*Regression*)
*JWHT: 59-70 (*Simple Linear Regression*)

Meeting 16: Interactions and Collinearity
AH: 34-36 (*3.3 Interactions*)
JWHT: 99-102 (*Collinearity*)

Lab 8: Intro to Linear Regression in R
Download the 'dogs' dataset on the course website.)

WEEK 9: Linear Regression Part II

Meeting 17: Transformations and Assumptions

W&S: 557-562 (*Assumptions of regression & Transformations*)

AH: 37-47 (*Statistical Inference & Graphical displays of data and fitted model*)

Meeting 18: Interpreting Results

AH: 53-74 (*Linear regression: before and after fitting the model*)

Understanding regression models and regression coefficients

(http://andrewgelman.com/2013/01/05/understanding-regression-models-and-regression-coefficients/)

Lab 9: Linear Regression in R: Interpreting results

Install the 'arm' package for 'display' function to interpret results)

WEEK 10: Logistic Regression Part I

Meeting 19: Intro to Logistic Regression

W&S: 563-575 (*Nonlinear regression*) & (*Logistic regression*)

AH: 79-85 (*Logistic Regression*)

Meeting 20: Building a Model

W&S: 605-625 (*Multiple explanatory variables*)

*AH: 85-104 (*Logistic Regression*)

Lab 10: Apply functions in R

Swirl: R Programming - 'lapply and sapply' and 'vapply and tapply'

WEEK 11: Logistic Regression Part II

Meeting 21: Generalized linear models

AH: 109-124 (*Generalized linear models*)

W&S: 593-596 (*Using species as data points*)

Meeting 22: Model Selection

AH: 524-526 (*Model comparison and deviance*)

JWHT: 175-181 (*Resampling methods*)

*W&S: 635-46 (*Computer-intensive methods*)

Lab 11 GLM Model building and interpretation of results

We will again use the 'dogs' dataset, which can be downloaded from the course website.)

WEEK 12: Brief Intro to Meta-analysis and Multilevel Models

Meeting 23: Power of Meta-analysis

W&S: 681-697 (*Meta-analysis combining information from multiple studies*)

MacLean, S.A & Beissinger, S.R. (2017).Species' traits as predictors of range shifts under contemporary

climate change: A review and meta-analysis. *Global Change Biology, 23,* 4094-4105.

Korner, C. (2017). When meta-analysis fails: A case about stomata.

Global Change Biology, 23(7), 2533-2534.

Meeting 24: Intro to Multilevel Models

AH: 237-248 (*Multilevel structures*)

AH: 251-262 (*Multilevel linear models: the basics*)

Lab 12 Cleaning large dataframes and multilevel model notation

Download the 'ospree' dataframe from the course website and make sure you read through

AH: 259-262 again before lab.)

Grading Rubric:

Type	**Percent of Grade**
Participation	10
Lab Problem Sets	20
Midterm	30
Final Exam and Project	40