IB 517 - The Analysis of Biological Data in R - Fall 2024

Instructor: Dr. Becky Fuller

Where: MWF 10:30-11:50 am 2233 Everitt Laboratory

Office Hours: 2-3 pm Mondays, 515 Morrill Hall

Contact Info: e-mail: rcfuller@illinois.edu
Lab Office: 104 Shelford Vivarium, 606 E. Healey
Office Hours: 515 Morrill Hall

Teaching Assistant: Ratna Karatgi

Contact Info: e-mail: karatgi2@illinois.edu
Office: 533 Morrill Hall
Office Hours: 3-5 pm,Fridays

Course Website: https://canvas.illinois.edu/courses/49136

Credit Hours: 4 hours

Prerequisite: a general course in statistics or consent of instructor.

Course Overview: The goals of this course are 2-fold. The first goal is to review and master commonly used statistical techniques including power analyses, chi-squared tests, t-tests, correlations, regression, ANOVA, non-parametric tests, generalized linear models, principal components analysis, model selection, and experimental design/interpretation. Most students will have had some prior exposure to these techniques. This course deals with these tools in-depth. At the end of the course, students should be able to use these techniques in their own research and should have a sufficient knowledge base that they can teach themselves related statistical techniques. The second goal is to get students comfortable programming in R. R has a 'steep learning curve'. Mastering R requires repeated practice with problem sets. At the end of the course, students should be able to create publication-quality graphs in R, analyze and simulate data, and trouble shoot their own code.

Required Textbooks: The Analysis of Biological Data 3rd edition, by Whitlock & Schluter; 2nd edition is fine

The Whitlock and Schluter book (W-S) is a book that you should own. It can be purchased as a hardback, paperback, or e-book.

Other Required Items: You should have access to a computer with a copy of R and R Studio loaded. Both are available as free downloads. We will discuss how to load these programs.

Student Learning Outcomes:

1. Students should thoroughly understand *how* basic statistical tools work. This includes being able to (1) calculate critical statistical parameters (i.e., means, differences in means, medians, correlations, standard deviations, and distributions) and (2) estimate the sampling distributions of those parameters.

2. Students should be able to correctly interpret statistics. This includes being able to simulate data to understand precisely what the parameters mean.

3. Students should be able to create graphics and tables that are of publication quality and easily interpretable.

4. Students should gain enough knowledge/tools/expertise on any one statistical subject that they can progress on their own after that.

Grades: This is a graded course. On the one hand, I really don't care about grades at the graduate level. Graduate school is all about what you can achieve and what you have to show for yourself at the end of your time here. On the other hand, you learn statistics and programming by ***doing*** them. By the end of this course, you should be able to teach yourself a lot of statistical tools.

Major Assignments: There are four major grade categories. There is a weekly reading quiz and a weekly computer lab that students must complete. The reading quiz will require you to read the Whitlock and Schluter book as well as the weekly discussion paper or interleaf. In the computer lab, you will graph, analyze, and interpret data using R. Students must also work in a group to lead a class discussion on a statistical topic and must participate each week in a discussion. Finally, there are three major independent projects that students must complete and present to their fellow students. These activities reflect our belief that students need to (a) come prepared to class, (b) repeatedly practice the tools that we cover, (c) engage in productive discussion with classmates, and (d) apply the tools we learn to questions that are meaningful to one's own research.

Here is the point breakdown:

Weekly Reading Quizzes - 30% (each reading quiz will count equally to the total)

Weekly Computer Lab Assignments - 30% (each computer lab assignment will count equally to the total)

Three independent projects = 30% (see specific instructions below the schedule)

a. Independent Project 1 – Find a poorly drawn graph from the published literature (preferably from your advisor). Explain how it violates the principles of good graphic design. Redraw the graph using R. Presentations will be made in class.

b. Independent Project 2 – Perform a power analysis of a study you would like to collect. Explain the question at hand, the statistical parameters to be estimated, and determine the sample sizes needed to detect a particular effect size. All simulations and graphics are to be done with R. Students will make a short video and submit online.

c. Independent Project 3 – Perform a linear model with two predictor variables (e.g., 2-way anova, mixed model, ancova, multiple linear regression, etc.). Obtain the data and analyze them in R. Check the assumptions of the analysis (i.e., normal residuals, no heteroscedasticity, etc.). Interpret the results. Presentations will be made during finals week.

Discussion of Articles/Interleaf - 10% (see specific instructions below the schedule)
Students will lead the discussion once with their classmates in a small group.
The group size will vary depending on the enrollment of the class. Students will also participate in the other discussions throughout the course led by their classmates.

Activity	Number	Points /Assignment	# Drops	Total Points	Percentage
Reading Quizzes	14	20	2	240	30%
Lab Activities	14	20	2	240	30%
Independent Projects	3	80	0	240	30%
Discussion	Lead Once + Participate	50 (lead) + participate(30)	2 times as participant	80	10%
			Total	800	

You may drop your two lowest pre-lecture quizzes and your two lowest computer lab assignments. Please note that I do this to accommodate student absences. I do not have a specific attendance policy, but I do expect students to be in class and lab. I understand that students get sick, attend conferences, do field work, and have time sensitive experiments. These drops are intended to cover these types of issues. Please do not ask me for an extension for these online quizzes. Of course, if you get extremely sick, then we can make arrangements.

The grading scale is as follows:
>90% of the points = A

89-80% of the points = B
79-70% of the points = C
<70% of the points = D or F

I will award plus and minus grades, but I will use my discretion to note the natural distribution of the scores. I also reserve the right to lower the grading scale (e.g., assign an A to a person who earns 89% of the points).

Schedule
I have posted the schedule of lecture topics, discussion papers, and due dates for projects.

I reserve the right to alter the schedule.

Computer Lab
There is a 1.5-hour computer lab for this course. We will lead students through analyses and give you data sets to graph, analyze, and interpret. We will also use this time to practice review problems from previous weeks.

Covid/Illness
Please do not come to class if you are (a) feeling ill, (b) have had a significant exposure to a person with covid, or (c) have to take care of small children when/if school is closed. We have videos on the course webpage that cover all of our topics. To be clear, I expect people to come to class if they are feeling well, but we can accommodate people working remotely. ***We will follow the CDC, Champaign County Public Health, and University of Illinois guidelines at all times.***

Academic Integrity:
Course policy on dishonesty follows Article 104 (1-401 through 1-406) of the Student Code, http://studentcode.illinois.edu/article1_part4_1-402.html, which should be consulted for details. To learn about likely penalties for violations of the Student Code on academic honesty, see http://studentcode.illinois.edu/article1_part4_1-403.html. It is your responsibility to know and understand the rules concerning academic honesty.

Your independent projects should be your own. You are encouraged to seek feedback from your fellow students, but you obviously have to do these yourself.

Please note that I am fine with students working on pre-lecture quizzes, lab assignments, and paper discussions in groups. However, **you may not post the answers to quizzes or lab assignments anywhere online.** This means no instagram, no facebook (even if a private group), no twitter/X, no quizlets, etc. Simply copying and pasting answers into quizzes defeats the purpose of working in groups. In addition, we will be completing many of the problems at the end of the chapters. Allowing the answers to these problems to be displayed online ruins the value of this book for other instructors who use it. **I will assign an F to anybody**

who posts answers to quizzes or homework assignments online. If I find out about such behavior after the course is done, I will retroactively assign an F.

Disability Accommodations

Disability Accommodations -To obtain disability-related academic adjustments and/or auxiliary aids, students with disabilities must contact the course instructor and the Disability Resources and Educational Services (DRES) as soon as possible. To contact DRES you may visit 1207 S. Oak St., Champaign, call 217-333-1970 (V/TTY), or e-mail a message to disability@illinois.edu.

Diversity Statement

Everybody has a right to study science statistics regardless of sex, sexual orientation, gender, nationality, race, or disability status. If you want to learn biostatistics and R, then I want to teach you. We will not tolerate people creating a hostile learning environment for others.

Class Schedule

Date	**Topic/Reading**	Reading Quiz?	Discussion Topic
	Week 1		
26-Aug	Lecture: Intro to the Course	yes	
28-Aug	Lecture: W-S: Chapter 1- Statistics & Samples		
30-Aug	Lab: Week 1: Working with the software, data structures, data manipulations		Cuddington et al. 2023, Building quantitative self-confidence in biologists
	Week 2		
2-Sep	**Labor Day**		
4-Sep	Lecture: W-S: Chapter 2- Displaying Data	yes	
6-Sep	Lecture: W-S: Chapter 2- Displaying Data		Wainer 1984. How to Display Data Badly.
	Week 3		
9-Sep	Lab: Week 2: Graphics - Part 1		
11-Sep	Lecture: Finish Graphing & W-S: Chapter 3- Describing Data	yes	
13-Sep	Lab: Week 3: Graphics - Part 2		Interleaf #1 Correlation does not require causation; #4 History of statistics

Week 4			
16-Sep	Lecture: W-S: Chapter 4- Estimating with Uncertainty	yes	
18-Sep	Lecture: W-S: Chapter 5- Probability		
20-Sep	Lab: Week 4: Basic Stats, Loops, Sample size & Uncertainty		Interleaf #2 Pseudoreplication, Colegrave & Ruxton, 2018
Week 5			
23-Sep	Lecture: W-S: Chapter 6 - Hypothesis Testing	Yes	
25-Sep	Lecture: W-S: Chapter 7 - Analyzing Proportions		
27-Sep	Lab: Week 5: Proportions & Hypothesis Testing		Interleaf #3 - Why Statistical Significance is not the same thing as Biological Significance & Duschoff et al. I can see clearly now.
Week 6			
30-Sep	Lecture: W-S: Chapter 8 – Fitting probability models to frequency data	yes	
2-Oct	Lecture: W-S: Chapter 9 - Contingency Analysis		
4-Oct	Lab: Week 7: Contingency Analysis, Frequency Data, Review		No Discussion Prepare for Presentation #1
Week 7			
7-Oct	**Lecture: Class Presentations - Project #1**	no	No Reading Quiz
9-Oct	**Lecture: Class Presentations - Project #1**		No Discussion
11-Oct	**Lecture: Class Presentations - Project #1**		No Lab Assignment
Week 8			
14-Oct	Lecture/Lab: W-S: Chapter 10 - The Normal Distribution; W-S: Chapter 11 - Inference for a normal population	Yes	Reading Quiz
16-Oct	Lecture: W-S: Chapter 12 - Comparing Two Means		
18-Oct	Lab: Week 8: The Normal Distribution & Comparing Two means		Interleaf #5 Making a Plan & Power Primer
Week 9			

21-Oct	Lecture: W-S: Chapter 14 - Designing Experiments, Power & Simulations	Yes	
23-Oct	Lecture: W-S: Chaper 13 - Handling Violations of Assumptions		
25-Oct	Lab: Week 9: experimental design and violations of assumptions, Power & Simulations		Interleaf #7 - Which test should I use? & Interleaf #6 - Controls in Medical Studies
Week 10			
28-Oct	Lecture: W-S: Chapter 15 - Comparing Means of More than Two Groups	Yes	
30-Oct	Lecture: W-S: Chapter 16 - Correlation		
1-Nov	Lab:T-tests and correlations		Interleaf #10 Publication Bias & #8 Data Dredging
Week 11			
4-Nov	Lecture: W-S: Chapter 17 - Regression Day 1	No	
6-Nov	Lecture: W-S: Chapter 17 - Regression Day 2		
8-Nov	**Lab: Week 11: Special Speakers, Regression Lab**		Interleaf #12 - Species as Data Points
Week 12			
11-Nov	Lecture: W-S: Chapter 18 - Multiple explanatory variables	Yes	
13-Nov	Lecture: W-S: Chapter 18 - Multiple explanatory variables		
15-Nov	Lab: Week 12: Regression, 2 factor anova		Interleaf #9 Experimental and statistical mistakes & Borer et al. 2009 Some simple guidellines for effective data management.
Week 13			
18-Nov	Lecture: W-S: Chapter 18 - Multiple explanatory variables	Yes	
20-Nov	Lecture: W-S: Chapter 18 - Multiple explanatory variables		
22-Nov	Lab: Week 13: Multiple explanatory variables (continued) **Presentation #2 due**		No discussion
Break Week			

25-Nov	**Thanksgiving Break**		
27-Nov	**Thanksgiving Break**		
29-Nov	**Thanksgiving Break**		
Week 14			
2-Dec	Lecture: Principal Components Analysis/Discriminant Function Analysis	Yes	
4-Dec	Lecture: W-S: Chapter 20 - Likelihood		Interleaf #11 Meta-analysis & Gurevitch et al. 2018
6-Dec	Lab: Week 14: Principal Components Analysis & Likelihood		
Week 15			
9-Dec	Lecture: W-S: Chapter 20 - Likelihood & Generalized Linear Models	yes	
11-Dec	Last Lab: Generalized Linear Models		Pragmatic approach to stats. Castilho & Prado 2021
Finals	**Class Presentations - Project #3**		

Independent Project #1: Graphics
5-7 Minute Presentation (Depending on Class Size)
Presentations on October 7, 9, 11

1. Find a published poorly drawn graph and obtain the underlying data. This may come from your advisor or a person in your lab. You may also choose a poorly drawn graph from the primary literature provided that it is at least 10 years old and has no accompanying Dryad/R file. If you choose this option, then you need to digitize the graph to get the original data. You can also simulate the original data if you have means, standard errors (or CLs), and a sample size.

2. Create a powerpoint presentation that shows the graph. Analyze the old graph. Explain the study and what patterns the graph is intended to display. Explain the flaws in the graph.

3. Redraw the graph in R using principles of effective display. Try to obtain and make use of the raw data, otherwise extract them from the graph or simulate them.

4. Analyze your new graph. Explain how your improvements display the patterns more effectively.

5. Include your R-code on a slide. Explain any parts of your code that you are particularly proud of.

6. Grade will be based on: the quality of your analysis of the bad graph; the degree of improvement of the new graph (choose your graph so as to leave yourself plenty of room for improvement); your interpretation of it; the quality of your R code. Please note that if you just make the graph 'pretty' (i.e., fix fonts, add color, etc.) then the maximum grade you can get is a B+. Graphs that re-draw the data in some way are eligible for an A.

Independent Project #2: Frequency Distributions, Simulations, and Power
Submit Videos By November 22

Think about a potential experiment that you might conduct that involves either frequency data (i.e., count data), a difference between two means, a proportion, a single mean, or a correlation between two variables.

Use a simulation to conduct a power analysis. Follow the instructions below. Make one slide for each bullet point below. You may add to the slides if you wish.

1. Slide 1: Introduce the topic; state why it is interesting.
2. Slide 2: State the scientific hypothesis that is being tested.
3. Slide 3: Explain the experimental design. Explain the nature of the replicates. How long will it take to perform the experiment?
4. Slide 4: What statistical test will you use and why? In other words, are you using a chi-squared test, a binomial test, a t-test between two independent groups, a paired t-test, a one-sample t-test, or a correlation? State the statistical null and alternative hypotheses being tested. Note that you are free to try more complex analysis if you wish. But this will take extra work on your part.
5. Slide 5: How will you simulate different levels of 'biological signal'? Are you investigating different scenarios that you anticipate a priori? Are you simulating different effect sizes? If so, what is your estimate of effect size (*r,* odds-ratio, cohen's D)? Are you testing whether a proportion (or a mean) differs from a null value? Again, how are you simulating different levels of biological signal? Explain the nature of your simulations.
6. Slide 6: Make a graph of a single 'run' of your experiment to show.
7. Slide 7: Simulate small, medium, and large effects (or particular types of biological signal). Is there a good justification for what the expected effect size might be? This slide is optional if you included this information in slide #5.
8. Slide 8: Show the results of the simulation. Use R to determine the # of simulations that did and did not result in a significant effect for a low, medium, and high sample size. From this, you will calculate the power. Use a graph to show the power (% significant) as a function of sample size and effect size. Interpret the graph.
9. Slide 9: Summarize your findings.
10. Slide 10: Show your R code. Feel free to brag about any part that you are particularly proud of.

Your presentation should be about 5-10 minutes. You will record this and upload your presentation. You can record on zoom, quicktime, etc.

Independent Project #3: Linear model

Presentations to be given during Finals. The finals schedule is still being determined. Some people may present earlier than finals.

You will create a PowerPoint presentation and present it to the class. Here are the instructions. Note that you may add extra slides if you wish.

- Obtain a data set from your supervisor or online data depository (e.g. Dryad).
- Slide 1: Explain the general purpose of the study that yielded the data.
- Slide 2: Explain the specific data set you are using. For example, say where the data are from, give the meaning of the variables, and so on.
- Slide 3: Show a graph of the data to illustrate the main patterns shown. Describe the pattern.
- Slide 4: State what parameters you will estimate with these data, e.g., the regression slope, differences among means of treatments, etc. State what you will test with these data.
- Slide 5: Fit a linear model to the data in R. Explain in words the model you fit.
- Slide 6: Interpret the output: Explain the test results. Show both the model summary and the ANOVA. If necessary, show the results of post hoc tests.
- Slide 7: Create a graph to visualize the model fit to the data. Explain what the graph is showing. In some cases, this will be similar to the graph that you showed in slide three.
- Slide 8: Address how well the statistical assumptions of your analysis were met, and how you handled violations.
- Slide 9: State the conclusions reached from your analyses.
- Slide 10: Include your clean R code on a slide.

  Please do not worry about whether or not your results are statistically significant. That is not the purpose of this analysis. The purpose is to analyze your data properly and thoroughly.

Some rules:

- Use only 1 response variable.
- Include at least 1 proper fixed factor, such as an experimental or observational treatment. Can be categorical or numeric.
- Include at least 1, and no more than 2, additional explanatory variables (random or fixed factors, blocks, covariates, etc).
- Try to keep sample size manageable to avoid computational issues (use subset if real data set too huge). Contact me if this might be an issue for your data set.

Weekly Discussions

Each week we will discuss either a published paper or one of the interleafs from the text. A small group of students will lead the class through the discussion. The exact group size will depend on the class size. Discussions last ~20 minutes. The rubric for the grading is below. There is a maximum of 50 points for leading the discussion.

Grading Rubric for Classroom Discussions

REQUIREMENT	APPROACHING MASTERY	ATTEMPTING MASTERY	NEEDS IMPROVEMENT
Introduction	Leaders included a thorough yet concise, one-minute overview of the topic/reading(s). (10 points)	Leaders included a one-minute overview of the topic/reading(s). (5 points)	Leaders did not include an overview of topic/reading(s) or the overview was too long. (1 point)
Questions	Leaders prepared a sufficient variety and number of open-ended, on-topic questions to stimulate a full 10-minute discussion. (20 points)	Leaders prepared on-topic questions that stimulated some discussion. (10 points)	Leaders did not prepare questions that stimulated much discussion and/or questions were off-topic. (6 points)
Facilitation and Participation	Leaders facilitated participation so that most classmates contributed to the discussion, and leaders' own participation was equal. (20 points)	Leaders facilitated participation so that many classmates contributed to the discussion and/or leaders' own participation was mostly equal. (10 points)	Leaders facilitated participation so that only some classmates contributed to the discussion and/or leaders' own participation was unequal. (6 points)
Total: X / 50 points Comments:			

Contribution to the group will be assessed by group peer review.

There is another 30 points available for participating in discussion. This grade is based on attendance and class participation.