

ESM 206 (Fall 2018)

Environmental Data Science & Statistics

Instructor: Allison Horst

Email: ahorst@bren.ucsb.edu

Usual office: Bren 4406

Office Hours: Monday 12:00pm – 1:30pm (Bren 3312)

TA: Jessica Couture

Email: jcouture@bren.ucsb.edu

Office Hours: Thursday 2:00pm – 3:30pm (Bren 3312)

TA: Sam Collie

Email: scollie@bren.ucsb.edu

Office Hours: Tuesday 1:00pm – 2:30pm (Bren 3312)

Course Description: Develop critical thinking, technical and communication skills to successfully approach and answer environmental questions using quantitative and qualitative data. Topics include: data wrangling, exploratory data analysis, descriptive statistics, uncertainty, hypothesis testing, data visualization and communication, and best practices for open data science, version control and collaboration. Skills will be developed through analysis of real-world datasets using R, RStudio and GitHub.

GRADING: 300 points total available in the course

Assignments (55%)

There will be 5 assignments, worth a **total** of 165 points, worth 55% of your total course grade.

Assignment 1: 27 points (9% of course grade)

Assignment 2: 30 points (10% of course grade)

Assignment 3: 36 points (12% of course grade)

Assignment 4: 36 points (12% of course grade)

Assignment 5: 36 points (12% of course grade)

Midterm (20%)

In-class midterm (in lecture Wednesday October 31st): 60 points (20% of course grade)

Final Exam (25%)

Take-home final (assigned December 5th, due December 11th): 75 points (25% of course grade)

IMPORTANT DATES:

Assignments:

- Assignment 1: Assigned Wednesday 10/3 – Due Monday 10/15
- Assignment 2: Assigned Monday 10/15 – Due Wednesday 10/24
- Assignment 3: Assigned Monday 10/29 – Due Wednesday 11/7
- Assignment 4: Assigned Wednesday 11/7 – Due Monday 11/19
- Assignment 5: Assigned Monday 11/26 – Due Wednesday 12/5

Exams:

- Midterm: In lecture Monday October 29th
- Final Exam: Take-home (December 5 – 11)

No-Class/Lab Dates (Week of 11/19)

- No lecture on Monday 11/12 (Veteran's Day)
- No lecture on Wednesday 11/21
- No labs on Tuesday or Wednesday (11/20 & 11/21)
- You **will** be given a take-home lab to work through on your own during the week of 11/19 – 11/23, which we expect will take ~3 hours

COURSE POLICIES:

- Assignments submitted late will only be accepted within one week of the due date, and will be worth 50% of the original score. **Homework submitted more than one week after the original due date will not be accepted.**
- Assignments are due at the beginning of lecture (9:30am) on due dates. Anything submitted after 9:35am on the due date is considered late and will be worth 50% of the original score. **You should expect Bren printer problems and traffic jams when considering this policy.**
- You are required to wait two full days before asking the instructor or TAs about scores on all returned course material
- Lab attendance is mandatory
- There will be no make-up assignments/midterms/quizzes or opportunities for extra points or extra credit due to low scores
- Cheating/plagiarism (including R code) will result in 0 points awarded for the assignment or midterm/quiz/exam and disciplinary action according to UCSB policy
- If you are worried about your overall grade and/or passing the class, it is **your** responsibility to reach out to us to discuss your concerns

TENTATIVE COURSE SCHEDULE AND TOPICS:

Week 1: Course introduction, problems and trends in open/data science, types of data, exploratory data analysis, intro to data visualization, accessing, loading and wrangling datasets in R using the tidyverse

Week 2: Exploratory data analysis (visualization, tables, scatterplots, summary statistics continued), introduction to Rmarkdown, probability density, Central Limit Theorem, exploring normality and data wrangling continued (sorting, subsetting, classes of data, histograms, QQ-plots, skewness and kurtosis)

Week 3: Uncertainty and bias, considering, calculating and expressing uncertainty in data analysis; probability continued (the z- and t-distribution); hypothesis testing; communicating uncertainty (graphics/tables/text), using GitHub for data science version control & collaboration (intro)

Week 4: Hypothesis testing continued, assumptions, null and alternative hypotheses, errors and considerations in hypothesis testing, limitations (preview: statistics isn't a p -value); effect size; power; communicating results of hypothesis tests (graphically, in tables and in text); hypothesis testing in R (performing, understanding outputs, communicating results); finding effect size and power (*a priori* power analysis for experimental design)

Week 5: Regression (ordinary least squares); assumptions, considerations, concerns/pitfalls; regression diagnostics in R; coefficient of determination, correlation (Pearson's r)

Week 6: Analysis of Variance (ANOVA), conceptually, mathematically and in RStudio; omnibus and post-hoc testing; communicating results. Working with count/categorical data; chi-square analysis, interpretation and communication; tables and counts in R; updating classes of data (factors/characters/integers) in R and specifying levels

Week 7: Multiple linear regression (conceptually, mathematically); interpreting multiple regression results; considerations in multivariate analyses; dummy variables and reference variables; pitfalls of the p -value; model diagnostics and concerns

Week 8: Multiple linear regression continued: exploring AIC, VIF, multicollinearity. What do we do when data doesn't satisfy assumptions of normality/parametric tests? Exploring alternatives: rank-based tests, non-parametric approaches, transformation; dealing with data limitations; types of bias in data collection, analysis and interpretation

Week 9: Rank-based tests continued (Wilcoxon Signed Rank, Mann-Whitney U, Kruskal-Wallis), pitfalls and considerations in rank-based tests

Week 10: Multi-factor ANOVA, non-parametric regression intro (binary logistic, ordinal logistic regression), moving forward from 206...