SC212: Introduction to Statistics and Data Science

Bret Zeldow

Spring 2022

E-mail: bmzeldow@colby.edu

Office Hours: See below or check Moodle

Office: Lovejoy 105 (located within the Lovejoy 101 suite)

Class Hours: MWF 9:00-9:50am

Class Room: Keyes 102

Course Description

> *The best thing about being a statistician is that you get to play in everyone's backyard.* – John Tukey

Statistics is the science of "making sense of data." Virtually all fields utilize statistics in some way. Not only that, but you are exposed to statistics practically every day — when you browse the internet, watch a sporting event, or listen to the news. Statistics are everywhere. This is true — even necessary — because data are everywhere, more so now than ever before. How can we use all these data to learn useful information? Extracting information requires more than learning a laundry list of methods; it requires careful consideration of context and the ability to think critically about data.

In this course, we will explore several tools that will help you make sense of data and draw conclusions. This includes methods for data collection, displaying and analyzing data, and developing and testing hypotheses. We will use RStudio, a desktop environment for the statistical programming language R, to explore real datasets and to analyze them. Over the course of the semester, you will learn how to use this software effectively and how to interpret its output. Additionally, we will spend considerable time learning and practicing how to communicate statistics effectively.

In this course, we will cover a lot of useful statistical methodologies such as regression, ANOVA, t-tests, but we will also cover many deeper ideas these methods. What is bias? What is confounding? What is a sampling distribution? Why should we care? When do traditional methods work? What should we do when they don't?

Course Objectives

1. To understand the methods of data collection and the importance of utilizing appropriate sampling methods and experimental design procedures.
2. To learn to recognize which data analysis and data visualization methods are appropriate for a given situation.
3. To learn basic data manipulation and data cleaning skills.
4. To understand the limits of statistical inference and understand the assumptions underlying common statistical methods.

5. To learn the implications of bias, variation, and confounding in the context of statistical inference as well as methods to address and control them.
6. To become comfortable with the use of statistical software and the basics of computational thinking.
7. To understand the critical role statistics plays in research applications across a variety of disciplines.
8. To demonstrate the ability to communicate statistical results effectively. This includes formulating scientific research questions, finding and evaluating peer-reviewed literature, referencing existing literature in reports, writing scientific reports that highlight data analyses, and presenting work in the form of a poster.

Primary Course Resources

Textbook

Statistics: Unlocking the Power of Data by Lock x5 — *Third Edition*.

Note on earlier editions: You are free to use earlier editions of the book with the understanding that the homework comes from the 3rd edition. No credit will be given to homework done with earlier editions of the textbook. The student is also 100% responsible for learning material that may appear in the 3rd edition but does not appear in earlier editions.

Online Supplement

While not required this semester, you may purchase a subscription to the WileyPlus online supplement for *Statistics: Unlocking the Power of Data*.

Software

We will use the R programming language with the RStudio desktop environment. You can access RStudio and R at http://rstudio.colby.edu. Or you can download R and Rstudio directly to your own machine.

We will also make use of STATKEY, which is a website companion to our textbook. STATKEY is a visual aid for understanding statistical concepts and a web-based tool that can perform statistical analyses.

Moodle

Assignments and announcements will all be accessible on Moodle. This is where I will post schedule changes and sometimes quizzes as well.

Topics Covered

Obtaining data/sampling
Observational studies and experiments
Visualizing and plotting data
Data manipulation and data cleaning
Linear regression
Hypothesis testing and confidence intervals
Bootstrapping and randomization tests
Density curves, normal distributions, and the Central Limit Theorem

Parametric hypothesis testing: t-tests and z-tests
Categorical data analysis: chi-square tests and Fisher's exact test
ANOVA: planned and unplanned comparisons
Multiple linear regression

Grading

Your grades will be determined in the following way:

	Weight
Assignments:	
Homework	10%
Labs + Moodle assignments/Worksheets	5%
Project	22%
Exams:	
Exam #1	21%
Exam #2	21%
Exam #3	21%

Grades will be recorded in Moodle and you can calculate your grade based on those and the weights given above. The overall grade seen in the Moodle gradebook is *NOT* your current grade as it doesn't take into account the above weights.

NOTE: To pass the course, students must average at least 60% on the three exams. Averages below 60% result in failing.

Lectures

Lectures will take place in-person from 8-850am in TBD.

Labs

Labs will provide hands-on experience with new datasets, the concepts we learn in class, and statistical software. Labs will take place in Davis 308 on either Wednesday or Thursday. Each lab will be scored with 0, 1, or 2 points. If your answers are complete and mostly correct, you will receive a 2. Labs must be turned in on Moodle by 5pm on Fridays. The lab work is only intended to take one class period (or maybe a little bit longer) to complete. **Turn your lab into your course professor,** *not* **your lab professor.**

Exams

Exams will take place during the usual class time on their scheduled days. If you have accommodations through the dean's office, I will happily grant those but *you* must contact me at least two weeks in advance of the exam.

Homework

Homework represents an important aspect of this course. Readings and selected problems will be assigned regularly. Assigned problems and due dates will be posted on Moodle. You are encouraged to discuss problems with other students, but each write-up should be your own. If you

produce the same computer output, that's okay. Just make sure the discussion and interpretation is in your own words. **Answers should be more than just a number; they should include thoughtful conclusions as well.**

Homework will consist of problems from the textbook. You may type or write your homework. In general, if you are producing graphics for the homework, they should be created in R and printed for the homework, rather than hand-drawn.

Homework will be submitted on Moodle as a single file in pdf format.

Delayed Exams/Projects

Students should treat deadlines as firm. Only valid medical excuses, documented personal catastrophes (such as a death in the family), and religious observances will be accepted as reasons not to take an exam or turn in an assignment as scheduled. If you are, for legitimate reasons, unable to take an exam or complete project deadlines when scheduled, you should notify me in advance of the exam time or due date. Having a lot of work to do, several exams/papers in a few days, being generally unprepared, or having conflicting travel arrangements are not acceptable excuses to miss an exam or assignment due date.

Late Assignment Policy (for labs, quizzes, and homeworks)

Homework

Each student will receive two free 48-hour extensions on homework (you may only use one per assignment) that you can use at any time, no questions asked. Lateness will be determined completely by the submission time on Moodle. Any assignment more than 48 hours late will be given a zero.

After the two 48-hour extensions are used, any late assignment is recorded as a zero.

Worksheets/Labs

Everyone will get *two* free skip for labs and worksheets/Moodle quizzes. If you don't use your skips, your lowest score will be dropped at the end of the semester. Please do not email me about skips — they will be applied automatically.

E-mail Policy

My policy is to respond to students' emails within 24 hours of receiving them during weekdays. You are always free to email me anytime, but please do not count on a response over weekends.

Office Hours

Office hours will be held in Lovejoy 105 (located within Lovejoy 101). Hours are posted below and are subject to change. Changes will be announced and posted to the Moodle page.

Day of Week	Times
Wednesday	2:15 - 3:30pm
Thursday	10 - 11am
Friday	10 - 11am