STAT 470 (Data analysis and statistical computation): Fall 2020

Instructor:	James Johndrow
	(OH:)
T.A.	
	(OH:)
Place and Time:	
First meeting:	

Work in progress: *This is my first year as instructor for 470, and I am in the process of updating the course. This syllabus is a work in progress but represents my best current picture of the course content and structure*

Agreement to abide by the syllabus: *By enrolling in the course, you acknowledge that you have read this document in its entirety, and agree to be bound by the course structure, grading rules, deadlines, and any and all other aspects of the course set forth herein. You also agree not to request changes in the structure of the course later that do not appear in this syllabus.*

Coronavirus note: *I have written this syllabus as though it is for an in-person course. Obviously, if part or all aspects of the course are taught online, some things will need to change (e.g. regarding assessment, note taking, et cetera). I will update this document as plans for the fall are clarified.*

1. Overview

1.1. **What is this course about?** This is a course on statistical computation. I mean this very literally: I will be teaching statistics, computation, and the interaction between the two. Specifically, we will be learning how to do statistics on a computer. This will naturally teach you some things that you need to know to do "data analysis." However, the term "data analysis" is so vague and nonspecific that you should really think of the course as "statistical computation."

The basic philosophy of the course is:

- Use of a computer language in general and R in particular can be hugely helpful in gaining a better understanding of the foundations of statistics
- The R language and statistical computing enable us to do statistics with fewer assumptions than more traditional methods that rely on calculating things analytically
- Learning to do statistics using a programming language like R also provides a practical basis for introducing fundamental concepts about computers and programming languages in general, a study that you could continue later by taking courses in computer science.

The main goals are

- Gain a thorough understanding of hypothetical repeated sampling, which is the bedrock principle of all of frequentist statistics, via computer simulation
- Learn a few basic concepts about programming languages and computing, like what algorithms, data structures, and storage types are.
- Learn a few of the most important computer-based tools in modern statistics: the bootstrap, permutation tests, Monte Carlo integration, and ways to correct for multiple testing.

1.2. **More details.** So what does learning about statistics and computation entail? Since this is a first course in this topic, and I do not expect that most of you have much exposure to programming, we'll be learning some basic things about programming languages and how they work, with a particular focus on things that you need to do statistics on a computer. These include things like algorithms, functions, iteration and flow control, and data structures. Computer science departments teach *entire courses* on things like algorithms and data structures, so necessarily what I give you will be very basic and just what you need to do the things we want to do in this course.

As far as the statistics content, I'm going to be teaching you how to use the computer to do things that you previously might have done mathematically, or perhaps by using "canned" software packages like JMP. Here you'll be doing a lot more "from scratch." In addition to needing to learn basic programming to do this – since R is a programming based not a point-and-click based environment – you'll also need to have a very good understading of the fundamentals of frequentist statistics. This means you'll need to understand how probabilities arise in statistics, and in particular how they are related to sampling and hypothetical replication of the sampling process. In reality you should already know this, but in practice you may find that your understanding of the very basic concepts is not as solid as you thought it was. As such, we will cover some of these basic concepts again. This won't be a repeat of what you have learned before, but rather will complement what I expect the focus of your previous exposure has been. For example, rather than teaching you once more how to compute a p value for tesing the hypothesis that the mean is zero in a normal model, we will talk about what a p value is *in general* and how we can use the computer to compute p values *in general*, then do examples. My goal is not to teach you 1,001 R commands. My goal is to teach you the fundamentals of how you can use R to answer questions and give you the tools to (begin) doing that in your academic or professional future.

1.3. **Prerequisites/expected background.** All of you should have had two semesters of statistics, either 101/102 or 111/112 (or, alternatively, 431). As such I expect that you know the basic foundations of frequentist statistics, such as

(1) Parameter estimation
(2) Hypothesis testing
(3) Interval estimation (confidence intervals)

You also should have had a course in mathematics that covers basic calculus (integration, differentiation/derivatives, ordinary differential equations). You cannot really take the

course without this background, you just won't know what I am talking about half the time.

1.4. **Course permissions.** I don't expect that most of you will need course permissions. In particular, I don't plan to waive the prerequisites, so please don't ask. I also don't expect that enrollment caps will be exceeded since we have 3 sections with room for a total of 225 students. If you are really, really sure that you need a permission and that this is not because you lack prerequisites, then please speak to me after class.

2. Course materials

There is no required textbook. I will provide background and instructional materials via lectures and code examples. Course related materials will be distributed via the Canvas website and scores for quizzes/homeworks/exams can be checked during the quarter using Canvas. Note that some of the course will be taught on the board, and that I may use handwritten notes. You should take notes when I write on the board. If you don't take any notes, and then later cannot remember what we discussed, this will fall under the heading of "not the professor's fault." If you miss class, get notes from a classmate.

You must install R and the Rstudio IDE. To install R, go here: `https://cloud.r-project.org/`. Select your platform, and download and install the latest version. **If you already have R installed, you still must download and install the latest version. This will save lots of time later. Also, if you do not download the latest version of R now, I will not help you with the problems that you are likely to encounter later in the course, and this may negatively impact your performance in the course**. To download Rstudio, go here: `https://rstudio.com/products/rstudio/download/` and download the free desktop version.

3. Assessment

Grades will be based on the following

(1) Homeworks: there will be 4 homeworks
(2) Quizzes: there will be 5 quizzes
(3) Midterm: there will be one midterm exam
(4) Final: there will be one final exam

The weighting will be: 40% final, 20% midterm, 20% homeworks, 20% quizzes. In general, all assessments are "cumulative" – anything covered in class is fair game. Please do not ask the question "what will be covered on [blank]?" because the answer is always "anything we have covered in class before then." You can drop the lowest quiz grade, but not the lowest homework grade. Late homeworks will be penalized by 10% for each day late, up to three days late, at which point you will automatically get a zero. These penalties are "off the top," not applied after the homework is scored. For example, if your homework is two days late, the highest possible score you can receive on that homework is an 80%. A homework will count as a day late beginning at 5pm on the day the homework is due. It will count

as two days late beginning at 5pm the day after, and so on. *Collaboration is allowed on homeworks, but not on any other graded material.* If you collaborate on homeworks you still must turn in your own homework. Homeworks will be submitted electronically on Canvas.

Quizzes and exams will be closed book/closed note. You may have a single piece of 8.5in by 11in paper with you for each quiz and exam. You can write anything you want on this, but you may not use magnifying implements (aside from eyeglasses or contacts) to read your notes.

4. Content

We have 14 weeks of instruction. This means I can teach you roughly 12 things, since it is quite hard to learn anything deep and worth learning in less than a week of instruction. The first week of the course will be review, an overview of where we are going, and some motivating types of problems that we will learn how to solve using R. The final week will be dedicated to synthesizing the material and reviewing for the final. This leaves us with the 12 weeks in the middle, during which I endeavor to cover the following:

(1) Fundamentals of computing I: data structures, assignment, memory, functions
(2) Fundamentals of computing II: programming basics, including flow control, iteration; more on functions
(3) Fundamentals of computing III: algorithms
(4) Estimation on a computer I: computing summary statistics, empirical distributions, empirical quantiles, et cetera
(5) Estimation on a computer II: least squares fitting, maximum likelihood, numerical optimization
(6) Prediction on a computer: regression models
(7) Introduction to simulation: random number generation, sampling variability, Monte Carlo
(8) Simulating variability I: the Bootstrap
(9) Simulating variability II: permutation tests
(10) Simulating variability III: additional topics
(11) Multiple testing and replicability: how to avoid being wrong most of the time
(12) Statistical learning: beyond simple regression models, cross-validation, sample splitting

5. Classroom expectations

Phones, laptops and other electronic devices are not to be used in class except for when we are actively using R in an interactive portion of the course (which will happen). You may use an iPad or other tablet at other times, but only if it is flat on the table.