STAT 237: BAYESIAN COMPUTING (SPRING 2022)

Contact and Logistics

Instructor. Colin Reimer Dawson (*they/them*)
Course Website. `http://colindawson.net/stat113/`
Course Slack Workspace. `https://stat237s2022.slack.com` (join via invite link on the website)
Class Zoom Room: (see email/Blackboard)
Hybrid Structure. We will have a mix of traditional "lecture" style meetings and hands-on "lab" meetings in this course, in roughly equal proportions. We may hold class on Zoom for the lab days, as I've found this format to be a bit more conducive to small group computer work than huddling around computers in a classroom, but we can decide this as a group, since it's not going to be the same days every week so the logistical hurdle of frequent alternations in format and the challenge of getting to and from other in-person classes to a Zoom setting may outweigh the advantages. If we do hold labs in person, you will need to have a laptop with a reliable internet connection.

Subject to Change Statement. The meeting modalities may change, either on an ongoing basis (if the college's policies change), or potentially for the occasional class here or there, depending on instructional objectives.

Office Hours.
Mondays 3:30-4:20pm (King 204, drop-in, but appts have priority)
Wednesdays 9:00-9:50am (King 204, by appt only)
Thursdays 11:15-1:00pm (King 203, casual group office hour/lunch/workspace)
Fridays 3:30-5:00pm (King 204, drop-in, but appts have priority)

General Learning Outcomes

By the end of this course you should be equipped to...

1. **reason about evidence and uncertainty** in informal contexts by combining background knowledge with new information
2. **translate basic scientific claims into a statistical model**

Date: Last Revised February 18, 2022.

3. **identify and reason about the unknown quantities (parameters) in such models**
4. **represent knowledge and uncertainty** about those parameters using probability
5. **perform basic manipulations of probabilities and probability distributions**
6. **use standard statistical software** to perform necessary computations to read in data and update our knowledge about unknowns
7. **communicate the findings and limitations** of a statistical analysis to an audience without specialized training in statistics

Course Overview

Historical Context: Bayesian vs "Frequentist" Statistics. Bayesian statistics is the name given to an approach to statistical reasoning named for the Rev. Thomas Bayes (1701-1761), who formulated a particular way of using probability to represent knowledge. Although Bayes is credited with the philosophical foundations, much of the mathematical formalization was worked out by Pierre Laplace (1749-1827).

The Bayesian approach differs in two main ways from the more common "frequentist" approach (associated primarily with 19th and 20th century scientists such as Francis Galton, Karl Pearson and Ronald Fisher) which is the focus of most traditional statistics courses (including STAT 113/114, 205, and 213 here at Oberlin).

The first and most often cited difference is that the Bayesian approach emphasizes the formal incorporation of context and background ("prior") knowledge into statstical inferences. The orignators of the frequentist approach rejected the formal use of prior knowledge in calculations, and attempted to keep background knowledge separate from the formal calculations as much as possible in the name of "objectivity". Background knowledge of course still guides the frequentist statistical process in the form of what data to collect, what models to consider and what interpretations to prefer, but it does not enter into the process of fitting models and estimating specific quantities.

The second difference, and the more important one in practice, in my view, is that the Bayesian approach represents uncertainty about the world directly using probability, whereas the frequentist approach treats probability as a tool for describing the behavior of processes and systems and not as a tool for knowledge representation. This means that a Bayesian analysis can assign quantiative degrees of plausibility to statements that may or may not be true, and treats decision making as a separate

problem from knowledge representation, tending to recommend that decisions be based on "expected utility" (how good or bad will the result of this decision be "on average"), whereas the frequentist approach focuses on controlling the frequency of incorrect conclusions in (some restricted version of) the "worst case" scenario.

The frequentist school dominated science for most of the 20th century after it was developed, partly because of its claim of greater objectivity, but mostly because applying the Bayesian approach to all but the simplest problems required solving complex integrals that are intractable on paper. However starting in around the 1980s, computational approaches to approximating these integrals involving iterative calculations were developed, and as computers became faster and more efficient versions of these algorithms were developed, it became practical to apply Bayesian statistics to a much wider range of statistical problems.

Our Agenda. The goal of this course is to provide an introduction to the Bayesian approach to statistical reasoning, develop basic facility with the mathematical machinery it uses (probability), and learn to use some of the modern computational tools that are required (chiefly an algorithm known as Markov Chain Monte Carlo, or MCMC) to perform the probability calculations required to draw conclusions. We will introduce the R programming language, accessed via the RStudio interactive development environment (IDE), and specifically two engines for MCMC (JAGS and Stan) that R has interfaces to (the `rjags` and `rstan` packages, respectively)

Who Should Take This Course. This course assumes no prior background in statistics, probability, or programming, though you will need a working knowledge of algebra and basic calculus for some calculations and to understand some concepts, and willingness to learn some programming (which can be a struggle if it's new to you, and possibly even if it isn't). The formal prerequisite is only MATH 133 or equivalent (Calculus I). This is a more mathematical course than STAT 113/114, and also relies more heavily on iterative algorithms performed by a computer, but it is equally introductory as far as statistical content. In essence it is a Bayesian counterpart to the frequentist introductions to statistics presented in STAT 113 and 114.

Structure of the Class

Meeting Format. Hands-on practice is absolutely essential to understand any formal system, and statistics is no exception. About half of our class meetings will consist of working in small groups on lab assignments, which will both reinforce content discussed in readings and other classes as well as introduce new concepts through guided exploration. The other half will be more conventional "lecture"-style meetings.

Graded Work. A focus on grades can get in the way of learning, as jumping through hoops needed to achieve a desired grade tends to impair deep thinking. That said, **as long as they take a back seat to a desire to genuinely understand and grow**, grades can sometimes provide a useful bit of concrete feedback, and external motivation when one's internal motivation flags (often due to busyness).

Honor Code. The Oberlin College Honor Code formalizes the idea that **all work that you submit is your own and that you have given credit to the ideas and work of others when you incorporate them**. You will be asked to write and sign the honor pledge on each graded assignment that you hand in. The honor pledge reads: "I have adhered to the Honor Code in this assignment."

What it means to adhere to the honor code depends on context. For each assignment type, I describe what it means to follow the honor code on that assignment below.

Homework. In addition to the in-class explorations, each of the labs will have 1-2 problems at the end for you to complete and turn in each week (some weeks the work due will include problems from more than one lab). I will also sometimes post additional problems for you to do for practice. These will not all be graded (and many will have solutions provided up front), but **it is important that you do them anyway** in order to stay on top of the content.

The graded homework and lab problems will be due electronically on Fridays by 5pm. I expect there will be about 11 homework assignments in total, so most weeks with a couple of exceptions. Each homework assignment will be graded out of 20 points.

You may freely collaborate with each other on homework problems and use any available sources, but this collaboration should be documented and each person should turn in their own writeup of the problems and code.

Takehome Exams. There will be one takehome exam on the preliminary/foundational content covered in the first three weeks of the semester (probability, and basic concepts of Bayesian inference), which will be due the Monday of Week 5 (3/21). This first exam will be written and mathematical content only – no computer component.

A second takehome exam will cover the core computational tools of Bayesian statistics and their application to common statistical models, which will be covered in Weeks 4-8. This exam will be due the Monday of Week 10 (5/02).

The exams will be graded out of 50 points each.

The exams need to be done individually, but you may use any sources that are either created by you (i.e., your notes) or which are assigned or linked to by me (i.e., the textbook and any supplementary source material I provide).

Final Project. There is no final exam. Instead you will conduct an end-to-end Bayesian analysis of a dataset and research question of your choosing, and write a paper outlining the problem context, and describing your models and methods. This can be done either individually or with one partner (up to you).

The final project will be graded out of 50 points.

Final Grade. At the end of the semester, the lowest homework grade will be dropped. In addition, you may revise and resubmit one of the exams at the end of the semester for regrading (you can and are encouraged to revise and resubmit both, but you will need to tell me which one to regrade). The resulting grade will be the average of the original grade and the revision grade.

Materials

Textbook. The main text is Doing Bayesian Data Analysis, 2 nd edition; by John Kruschke.

Software. We will use the free statistical computing environment RStudio, which is an interface to the free and open-source language R. Once I set up an account for you, you can access the software via Oberlin's RStudio server via a web browser (`rstudio.oberlin.edu`).

The software is free, and you can optionally install R and RStudio on your personal machine (`www.r-project.org` and `www.rstudio.com`, respectively); however, this will require a bit more management on your part, and you will still need to log in to the server version to submit assignments and access feedback and solution sets, so unless you have some prior experience with computing I recommend sticking with the browser interface for now to minimize confusion.

Miscellany

Communication Outside Class Time. I have set up a Slack workspace for communication related to the course. You can join via the link posted on the course website (requires an oberlin.edu email address). **I am likely to respond more quickly if you message me there rather than via email**; however, don't hesitate to follow up if you don't hear from me within a day or two, as sometimes things slip through the cracks.

If you have a question or comment that other students might be interested in, I encourage you to **post to one of the classwide channels** rather than PMing me. You might even get a faster response from one of your peers than from me!

If you need to ask me about something due the following morning, don't wait until the night before! I have family and parenting responsibilities in the evenings and on weekends, and cannot necessarily respond to messages outside normal "business hours".

Accommodations. I have tried to structure the course in a way that automatically accommodates the most common situations; there are no timed in-class assessments, for example. However if you require other accommodations to do your best work in this class, please let me know as early as possible, and consult as well with the Office of Disability Services (ODS). By college policy, **all requests for accommodation require documentation from ODS.**

Major Dates and Times

Fridays 5:00 PM	Homeworks Due Electronically
Monday 03/21, 11:59 PM	Takehome Exam 1 Due
Monday 05/02, 11:59 PM	Takehome Exam 2 Due
Friday 06/03, 9:00 PM	Final Project Writeup Due

Schedule of Topics

See the "Schedule" tab on the course website (`http://colindawson.net/stat237/schedule/`).