What you will learn

- **Bayesian data analysis** through probabilistic programming (R/Stan, Python/PyMC, Julia/Turing)\(^1\)
- Computational approaches for statistical modeling (quadratic approximation, sampling, MCMC, Hamiltonian Monte Carlo)\(^2\)
- Causal inference to develop robust statistical models and identify causal relationships
- Machine learning techniques to reduce overfitting in statistical models (e.g., cross-validation, regularization, shrinkage, pooling)

Course overview

This course builds your knowledge of and confidence in making inferences from data. Reflecting the need for scripting in today’s model-based statistics, students will perform step-by-step calculations that are usually automated. This unique computational approach ensures that sufficient understanding to make reasonable choices and interpretations in your own modeling work.

We’ll cover causal inference and generalized linear multilevel models from a simple Bayesian perspective that builds on information theory and maximum entropy. The core material ranges from the basics of regression to advanced multilevel models. If time, we’ll discuss measurement error, missing data, and Gaussian process models for spatial and phylogenetic confounding.

The course also includes directed acyclic graph (DAG) approach to causal inference. Additional topics may include the design of prior distributions, splines, ordered categorical predictors, social relations models, cross-validation, importance sampling, instrumental variables, and Hamiltonian Monte Carlo. Our goal is to go beyond generalized linear modeling, showing how domain-specific scientific models can be built into statistical analyses.

Course philosophy

Classical statistics classes spend substantial time covering probability theory, null hypothesis testing, and other statistical tests first developed hundreds of years ago. Some classes don’t use software or actual real data and instead live in the world of mathematical proofs. They can be math-heavy and full of often unintuitive concepts and equations.

In this class\(^3\) we will take the opposite approach. We begin with data and learn how to use programming to draw inferences from statistical models.

In other words, there’s way less of this:

```
P(W, L | p) = \frac{(W + L)!}{W!L!} p^W (1-p)^L
```

And way more of this:

```
dbinom(6, 9, 0.7)
```

Over the last decade there has been a revolution in statistical and scientific computing. Open source languages like R and Python have overtaken older (and expensive!) corporate software packages like SAS and SPSS, and there are now thousands of books and blog posts and other online resources with excellent tutorials about how to analyze pretty much any kind of data.

This class will expose you to R—one of the most popular, sought-after, and in-demand statistical programming languages. Armed with the foundation of R skills you’ll learn in this class, you’ll know enough to be able to find how to analyze any sort of data-based question in the future.

Important pep talk!

I promise you can succeed in this class.

Learning Bayesian Statistics and R for any programming language can be difficult at first—it’s like learning a new language just like Spanish, French, or Chinese. Hadley Wickham—the chief data scientist at RStudio and the author of some amazing R packages you’ll be using like ggplot2—made this wise observation:

---

\(^1\) Dr. Ryan Wesslen, rwesslen@uncc.edu
\(^2\) @ryanwesslen
\(^3\) Center City 714

---

https://dsba6010-spring2022.netlify.app/syllabus/
It's easy when you start out programming to get really frustrated and think, "Oh it's me, I'm really stupid," or "I'm not made out to program." But, that is absolutely not the case. Everyone gets frustrated. I still get frustrated occasionally when writing R code. It's just a natural part of programming. So, it happens to everyone and gets less and less over time. Don't blame yourself. Just take a break, do something fun, and then come back and try again later. Even experienced programmers and evaluators find themselves bashing their heads against seemingly intractable errors. If you're finding yourself taking way too long hitting your head against a wall and not understanding, take a break, talk to classmates, e-mail me, etc.

Learn to draw the owl

**H ow T o D ra w A n O w l**

1. Draw some circles
2. Draw the rest of the owl

Learn how to learn effectively

Be confident to be comfortable with confusion
Benefits from taking this course

- Develop unique skill set (Bayesian statistics + causal inference + probabilistic programming + ML) that will distinguish you in the job market and/or academic opportunities.
- Use problem sets and the final project to build your data science portfolio (encourage communication and reproducibility).
- Exposure to a wide range of applications in Bayesian inference and causal inference in areas like astronomy, ecology, human-computer interaction, cognitive science, biology, political science, and many more.
- Exposure (when possible) to additional data science tools/techniques to enhance how you communicate data analysis (e.g., RMarkdown / quarto, shiny / streamlit, tidyverse).

Course materials

Book and materials

There is one textbook for this course: Richard McElreath’s *Statistical Rethinking*. It is very important that you purchase the 2nd edition. Simultaneously this semester, Richard is teaching an online version of his course. We will leverage his online lectures, slides, and problem sets for our course. Richard has taught the course many times and his iterations has led to a course (Winter 2019 and Winter 2020) much better than anything I could create.

I highly recommend viewing his book website that includes textbook code, an tidyverse version of the code, PyMC (Python) version of the code, and turing.jl (Julia) version of the code.

R and RStudio

You will do all of your analysis with the open source (and free!) programming language R. You will use RStudio as the main program to access R. Think of R as an engine and RStudio as a car dashboard—R handles all the calculations and the actual statistics, while RStudio provides a nice interface for running R code.

R is free, but it can sometimes be a pain to install and configure. To help you, I've created (and will likely add additional) guides in the Resources section of the website. We'll cover these in the first class.

Can I use Python or Julia instead of R in the class?

Online help

Data science and statistical programming can be difficult. Computers are stupid and little errors in your code can cause hours of headache (even if you’ve been doing this stuff for years!)

Fortunately there are tons of online resources to help you with this. Two of the most important are StackOverflow (a Q&A site with hundreds of thousands of answers to all sorts of programming questions) and RStudio Community (a forum specifically designed for people using RStudio).

If you use Twitter, post R-related questions and content with #rstats. The community there is exceptionally generous and helpful.

Searching for help with R on Google can sometimes be tricky because the program name is, um, a single letter. Google is generally smart enough to figure out what you mean when you search for “r scatterplot”, but if it does struggle, try searching for “rstats” instead (e.g. “rstats scatterplot”).

Additionally, we have a class chatroom at Slack where anyone in the class can ask questions and anyone can answer. I will monitor Slack regularly and will respond quickly. Ask questions about the readings, assignments, and project. You’ll likely have similar questions as your peers, and you’ll likely be able to answer other peoples’ questions too.

Course structure

We meet weekly from 12:00-2:45 PM on Mondays in Center City 1101. This course will employ a flipped classroom teaching method.

We will not have lectures during our regularly scheduled class time. Instead, you will do the readings and watch recorded lecture videos prior to each in-person class session. You can do the readings and watch the videos on your own schedule at whatever time works best for you. We will do several things during our Monday in-person classes:
• **Extensive Q&A**: As you do the readings and watch the videos prior to class, you will inevitably have questions. In your weekly check-in, you can submit (at least) 3 of those questions to me prior to class. We’ll spend a good chunk of each class answering, clarifying, debating, and discussing your questions. We’ll also review the lecture quiz due at the beginning of the class to ensure students are familiar with the key concepts of the material.

• **Activities**: In some weeks, we’ll do some in-class activities labeled as examples to help solidify concepts about Bayesian statistics, causal inference, and programming. These will not be graded but may aid in upcoming problem set.

• **Problem sets**: We’ll spend a substantial time during each class learning and working with R together on the problem sets. You’ll need to bring a computer.

### Schedule

The course schedule is listed here. In general, we’ll likely cover 1-2 of Richard’s lectures per course. However, there will be a few exceptions. This schedule may change per the instructor’s discretion.

### Weekly check-in

Every week, after you finish working through the content, I want to hear about what you learned and what questions you still have. Because the content in this course is flipped, these questions are crucial for our weekly in-class discussions. To encourage engagement with the course content—and to allow me to collect the class’s questions each week—you’ll need to fill out a short response on Google Forms.

### Assignments and grades

There are five components to your grade in the course.

1. **Participation**: Students will receive 5% of their grade based on participation. While I will not actively take attendance, this will be in large part a function of in-class participation (e.g., asking/answering questions, weekly check-ins) as well as out-of-class (e.g., Slack).

2. **Lecture quizzes**: Before each class (due 11:59am of each class), students will complete a 5 question quiz on the due lecture/chapter. There will be a 10 minute timer and the quizzes will use Canvas. There are no late quizzes that will be accepted as we will review the quizzes first thing for each class. There are eleven (11) total quizzes; however, the two (2) lowest quizzes will be dropped at the end of the semester.

3. **Problem sets**: For each lecture, there will be an accompanying problem set. These problem sets will largely mimic those from Richard’s classes with some additional problems. The lowest problem set will be dropped. Problem sets late will receive a 50% reduction. Please take note of the problem set grading rubrics.

4. **Exam**: There will be one in-class exam. It will be closed notes and based on chapter lectures, slides, and materials. It will be a blend of multiple choice (very similar to lecture quizzes) and short open ended questions (e.g., reading comprehension or explaining code snippets similar to problem sets).

5. **Final Project**: The final project will provide students an opportunity to find an application for Bayesian inference. Part of the grade will be two check-in assignments including an in-class presentation on a Bayesian topic and a proposal assignment.

You can find descriptions for all the assignments on the assignments page.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Points</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>50</td>
<td>5.0%</td>
</tr>
<tr>
<td>Lecture quizzes</td>
<td>180</td>
<td>18.0%</td>
</tr>
<tr>
<td>Problem sets</td>
<td>320</td>
<td>32.0%</td>
</tr>
<tr>
<td>Exam</td>
<td>200</td>
<td>20.0%</td>
</tr>
<tr>
<td>Final Project</td>
<td>250</td>
<td>25.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1000</strong></td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Comendable)</td>
<td>900-1000</td>
</tr>
<tr>
<td>B (Satisfactory)</td>
<td>800-899</td>
</tr>
<tr>
<td>C (Marginal)</td>
<td>700-799</td>
</tr>
<tr>
<td>U (Unsatisfactory)</td>
<td>&lt; 700</td>
</tr>
</tbody>
</table>

▶ Can I work with others for problem sets?

### Classroom policies

#### Masks

Per UNC Charlotte’s indoor mask policy:

https://dsba6010-spring2022.netlify.app/syllabus/
Until further notice, face coverings are required in all indoor spaces at UNC Charlotte and are strongly encouraged outdoors when physical distance cannot be maintained. Face coverings must fully cover the nose and mouth.

This requirement is for all individuals regardless of vaccination status. It applies to all spaces, including Atkins Library, research spaces and studios, dining halls, recreational facilities, common spaces and residence halls. There will be only rare exceptions to this requirement, such as when students are in their personal residence hall rooms or when employees are in their personal offices.

Failure to comply with this requirement may result in disciplinary action under the Code of Student Responsibility or employee personnel action.

Vaccines
If you are not vaccinated (with booster), please consider getting the COVID-19 vaccination. It is free. It saves lives. (Full disclosure: I am fully vaccinated plus booster)

Orderly and productive classroom conduct
I will conduct this class in an atmosphere of mutual respect. I encourage your active participation in class discussions. Each of us may have strongly differing opinions on the various topics of class discussions. The conflict of ideas is encouraged and welcome. The orderly questioning of the ideas of others, including mine, is similarly welcome. However, I will exercise my responsibility to manage the discussions so that ideas and argument can proceed in an orderly fashion. You should expect that if your conduct during class discussions seriously disrupts the atmosphere of mutual respect I expect in this class, you will not be permitted to participate further.

Recording in the classroom
Electronic video and/or audio recording is not permitted during class unless the student obtains permission from the instructor. If permission is granted, any distribution of the recording is prohibited. Students with specific electronic recording accommodations authorized by the Office of Disability Services do not require instructor permission; however, the instructor must be notified of any such accommodation prior to recording. Any distribution of such recordings is prohibited.

Discussion of grades and performance
Such discussion shall occur between the student and the instructor(s). Sharing information regarding grades and performance in places such as discussion forums or email blasts is prohibited.

Code of Student Responsibility
“The purpose of the Code of Student Responsibility (the Code) is to protect the campus community and to maintain an environment conducive to learning. University rules for student conduct are discussed in detail. The procedures followed for any Student, Student Organization or Group charged with a violation of the Code, including the right to a hearing before a Hearing Panel or Administrative Hearing Officer, are fully described.” (Introductory statement from the UNC Charlotte brochure about the Code of Student Responsibility). The entire document may be found at this site: https://legal.uncc.edu/policies/up-406

Academic Integrity
All students are required to read and abide by the Code of Student Academic Integrity. Violations of the Code of Student Academic Integrity, including plagiarism, will result in disciplinary action as provided in the Code. Students are expected to submit their own work, either as individuals or contributors to a group assignment. Definitions and examples of plagiarism and other violations are set forth in the Code. The Code is available from the Dean of Students Office or online at: https://legal.uncc.edu/policies/up-487.

Faculty may ask students to produce identification at examinations and may require students to demonstrate that graded assignments completed outside of class are their own work.

FAQs
- What are the course prerequisites?
  1. https://twitter.com/Chelseaparlett/status/1202152930359587017?s=20
  2. https://twitter.com/Chelseaparlett/status/1250767616492265779
  3. This philosophy as well as many other inspirational resources were originally created by Andrew Heiss, for example: his PMAP 8521 course website. Therefore, he deserves all praises for these ideas (he’s an amazing teacher). I’m grateful to be able to use such materials.
  4. I highly recommend aspiring data scientists to read David Robinson’s inspiring post to create a blog. If you’re interested for yourself, I also recommend starting with Allison Hill’s Up and Running with blogdown post to show how to do it in R. In fact, this website is created using blogdown. DM me on Slack if you have questions.
  5. My philosophy on problem sets is motivated by tweets from Cora Wigger to find a right balance between enabling students to keep pace (provide answer key for some questions) while providing an incentive to work on the problem sets (some questions w/o answer key).
  6. Note: this exam structure may change if we need to move to an online due to the ongoing pandemic. Details will be provided in advance to the exam.
  7. UNCC does not require this and I can’t legally require this but I am allowed to urge it so here’s me urging it. Thank you for understanding and protecting your classmates’ health.

DSBA 6010 | STAT 7027: Bayesian Statistics (Spring 2022)
UNC Charlotte School of Data Science Dept. of Mathematics and Statistics

All content licensed under a Creative Commons Attribution-NonCommercial 4.0 International License
Content ©2022 Ryan Wesslen

https://dsba6010-spring2022.netlify.app/syllabus/