STSCI 4780/5780 - Bayesian data analysis: principles and practice

For simplicity, STSCI 4780/5780 is often referred to as **STSCI 4780** or **BDA** in course documents and correspondence.

Lectures: Tuesdays & Thursdays, 1:00pm - 2:15pm, in PhySci 120
Labs: Fridays, 2:45pm - 4:00pm, in Warren B75

Instructor:
Tom Loredo
Senior Research Associate, Cornell Center for Astrophysics and Planetary Science
Field Faculty Member and Lecturer, Graduate Field of Statistics
620 Space Sciences Building
loredo@astro.cornell.edu
Office hours: Wednesdays, 2:30pm - 4pm, and by appointment (please avoid Mondays)

Teaching Assistant:
Georgia Smits: ges256@cornell.edu
Office hours: TBD

Class discussion forum: An Ed forum is accessible via the course's Canvas site. We've previously used Piazza for online class discussions; please bear with us as we adapt to Ed.

Contents

Enrollment waiting list

Enrollment for STSCI 4780 is currently full. **If you are hoping to add the course if spaces open up, please email Tom** (if you haven't already), so you can get access to course announcements and material in private GitHub repositories. Students waiting for an opening must monitor the course enrollment themselves to see if an opening arises.

Course goals

Provide students:

- A basic understanding of the principles and foundations underlying the Bayesian approach
- Practical experience using basic/intermediate Bayesian methods
- Experience with some widely-used tools and software development practices for producing and sharing collaborative, reproducible statistical research
- Exposure to the Bayesian academic research literature
- An understanding of key differences between Bayesian and frequentist approaches

Grading & assignments

Grading will be based almost entirely on homework assignments (current plan). Note that assignments will have varying difficulty, and thus contribute different amounts to your grade. The final assignment will be a challenging two-week assignment.

All assignments will be in the form of *Jupyter notebooks*, submitted for grading via student Git repositories in the course's GitHub organization. The early lab sessions will provide an introduction to Git, GitHub, and Jupyter notebooks. Newcomers to this technology may want to seek additional tutorial content, e.g., via video courses accessible via Cornell's subscription to the LinkedIn Learning platform.

Provide ***solutions****, not answers.* Communicate your *reasoning*, not just your result. A solution with sound reasoning but a minor error (like an innocuous sign error) will get more credit than a correct answer presented without motivation or with incorrect reasoning.

Everyone has a rough week now and then. The assignment that has the most negative effect on your grade will be dropped in everyone's final grade calculation (except that the final assignment will not be dropped). If you wish, you may skip an assignment without prejudice, but please do so cautiously.

Hand in assignments on time. New assignments will be provided during the Friday labs, and labs will typically run through material relevant to a new assignment. Completed assignments will be due shortly before the next lab. Especially starting around Assignment04, **many assignments will build on each other**. For such assignments, new assignments will contain code that corresponds to the solutions of the previous assignment. It is thus extremely problematic to hand in assignments late.

If you need an extension, request one as soon as possible; don't wait until the last minute. We will readily grant extensions early in the course (largely because we understand that some of you are starting the course less well prepared than others, particularly in regard to the technology we're using—Python, Git, GitHub, Markdown, MathJax). After the early assignments, please, ask for an extension *only under unusual circumstances*. Illness, multiple job interviews on the assignment due date, a significant family or personal matter—these are all circumstances that may justify an extension. Your course load generally is *not* a good reason for an extension after the early assignments—plan ahead to ensure you can complete assignments on time.

Class participation matters. As statisticians, clear communication of understanding and uncertainty is something that will be expected of you, and you need to be able to do this verbally as well as in documents. I'll be keeping track of participation (questions and answers, in class, labs, and outside-of-class discussion, e.g., in the class forum and office hours). Ask questions when you are puzzled; don't think your question is "dumb" (chances are other students have the same question—perhaps because I made an error!). Answer queries and questions boldly; I don't really care whether you give the "right" or "wrong" answer to a question (indeed, many questions I pose won't have a unique right answer); I mainly want to see you genuinely engaged with the material, and with the class. Although class participation will not formally enter the grade calculation, for students at a boundary between grades, participation will be a factor influencing the assignment of the final grade.

Grading will be on a curve. That said, in all past semesters, the grade boundaries (as a percentage of the total possible score on assignments) have been close to Cornell's past standard *un*curved grade boundaries:

```
96–100 = A+
93–95 = A
90–92 = A
86–89 = B+
83–85 = B
80–82 = B–
76–79 = C+
73–75 = C
70–72 = C
66–69 = D+
63–65 = D
60–62 = D–
59–lower = F
```

Per university policy, students registered for an S/U grade will get an S grade if their (curved) letter grade would be C- or higher.

Collaboration

I learned this material long ago. I'm probably too familiar with some of it to know how best to explain it to every person who is new to it. So some of you may learn best by talking over lecture or assignment content with your classmates, who have just figured it out for themselves, perhaps with an approach that appeals more to you. Thus, as a general rule, I encourage you to collaborate with each other, both with questions about the lecture material, and with issues that come up in the assignments (which I try to structure to be significant learning exercises).

However, ***don't simply copy another student's work***. You may *discuss* assignments with each other, and even look at each others' code or derivations, but you must ***do the work yourself***, writing your own solution or code, in your own words/style. There may be some assignments where I want you to work on your own to a greater extent (esp. the final assignment); this will be specified in the assignment instructions.

More specifically, here are example collaboration scenarios that are forbidden and allowed:

- You may *not* share solution code or derivations (yours or someone else's, including from a previous year, or from an online source) by email, on the forum, in writing, or via any other method that enables someone to easily or directly copy solution code. You may share *non-solution code* through such channels—e.g., to discuss code provided in an assignment, or code in the documentation for Python or Python packages or tools—but not code comprising all or part of a solution.
- Using someone else's code, but altering variable names, indentation, or other ancillary details will be considered as equivalent to verbatim copying, and comprises a serious academic integrity violation.
- You may work together with other students, in person, or virtually (e.g., via a Zoom conference). You may look at each others' work by viewing each others' screens (in person or via screen sharing). In the case of virtual screen sharing, you must not take a screen shot or make any other persistant copy of another student's work. Such sharing should be used only to build understanding that can help you come up with *your own* effective solution to a problem.
- An ideal way to collaborate, when you are having trouble with a problem, is to have another student (who may be having more success with the problem) look at *your* code (rather than you look at theirs), to give you advice on what could be improved. If you are an advising student in such a situation, try not to simply give the answer; rather, point out possible issues you may see, and direct your collaborator in how to think about the problem (e.g., by pointing them to relevant lecture content or material from another source, or by pointing out something that may be wrong, allowing your collaborator to work out what should replace it).

If you are uncertain about whether a particular collaboration scenario may be allowed or not, contact the instructors about it.

Collaboration disclosure. If you have collaborated closely with one or more classmates (either giving or receiving help or advice), or drawn significant material from an online source, book, or other resource, you should disclose the nature of the collaboration/consultation in your solutions notebook, in a single *collaboration disclosure cell* near the top of your solutions notebook. The disclosure should identify your collaborators, and *briefly* (in just a sentence or two) describe the nature of the collaboration (including if you helped others). E.g.,

- Jane Doe helped me debug my code in problems 2 and 3.
- I [Jane Doe] helped Sam debug code for a couple problems.
- Jack Smith helped me when I got stuck with the algebra in problem 1.
- I worked on and off with Teresa and Xiulin and we discussed problems when we got stuck. [This implies more or less equal help across the group, that wasn't too detailed.]

We don't anticipate giving detailed attention to disclosures (e.g., we aren't going to be too picky about consistency, e.g., if you forget to mention you helped someone who says you helped them in their disclosure). The disclosures are meant to help us understand what's going on when solutions look a bit similar. Mainly we hope they'll encourage you to try to make your work as independent as possible, while still benefitting from collaboration, and to give each other a shout-out for helping each other learn the material.

Collaboration via the discussion forum. You may also use the course's discussion forum to ask (and answer!) questions about lectures, labs, and assignments.

View the forum primarily as a virtual study group attended by all the students in the class, and only secondarily as a way to ask questions of instructors. Consider fellow students to be the primary audience for forum posts. In that light, we encourage you to post any interesting resource, insights, or news items relevant to the course on the forum, not just questions about assignments. We consider forumposts—especially those that contribute information, including answers to questions—as a form of class participation.

We'll monitor and contribute to the forum as best as we can, especially for questions seeking *quick and brief* clarification on a homework problem or lecture topic. But if you need nontrivial help specifically from an instructor (lecturer or TA), attending office hours (regular or ad hoc) is the main way you should seek such help. In particular, do not ask for help on the forum with the expectation of getting a quick instructor response outside of regular work hours, especially on the evening an assignment is due. We may respond on the forum at such times, but you should not count on it.

The key to getting good help on the forum (from instructors or your peers) is to *ask good questions*. Michael Clarkson, in Cornell's CS department, provides advice on writing good forum questions here: Asking Technical Questions - CS 3110 Fall 2019. There are many good tips in this document, even on how to title your question to help it get more attention from fellow students.

Keep in mind that *the forum is a very poor channel to use for jointly debugging code*. In most cases, fixing a bug requires exploring parts of the failing code up the calling chain from where the bug becomes apparent. Trying to do this by iterating on the forum is extremely frustrating and time consuming. And in any case, you must not share solution code publicly on the forum. A *private* forum question, to instructors, is also not appropriate for debugging code (private or public posts with conceptual questions about homework problems are fine, including posts with plots from your solution). If you need debugging help, try to attend office hours or schedule a one-on-one Zoom session, so you can share your code editor screen with an instructor and iterate more quickly toward a solution.

Note that the forum may permit anonymous posting. If that's the case, the forum will likely be set up so that an anonymous post reveals your identity to instructors, but hides it from fellow students.

Academic well-being

College/university education can sometimes be stressful; the additional burdens added by pandemic impacts and remote instruction can significantly add to the stress. Cornell provides resources for students to help promote academic well-being and good mental health:

- Cornell's Learning Strategies Center (LSC) provides tips for more efficient, less stressful learning. On the topic of collaboration, they provide a service to help students find study partners; see: Studying Together. If you are seeking a study partner and LSC can't help, contact me about it.
- Regarding online instruction, LSC also has resources to help you learn more effectively online; see: LEARNING ONLINE – Learning Strategies Center.
- LSC operates *Motivation Stations* for students who can benefit from learning in the presence of peers—a kind of Zoom-based study hall, with tutors available to help with study skills. See: Motivation Station and Study Skills Peer Consults! – Learning Strategies Center.
- LSC has created Canvas modules on various study skills: "Gearing up for Academic Success", "Managing Space and Time", "Taking Effective Notes", and "Preparing for and Taking Exams". To enroll: Enroll in LSC Study Skills Modules.
- Other Cornell-based resources that can help with mental health and stress issues include:
  - Skorton Center for Health Initiatives — Alcohol and other drug initiatives, anti-racism and bias prevention, hazing prevention, mental health promotion, sexual violence prevention, suicide prevention
  - Cornell Health
  - Mental Health at Cornell, including a directory of 24/7 help resources: 24/7 Help
  - Student-led mental health initiatives: Ways Students Can Get Involved | Mental Health at Cornell

Academic integrity

Academic integrity expectations for STSCI 4780 (beyond those addressed above in regard to collaboration) are spelled out in a separate document: Academic integrity expectations for STSCI 4780.

Instructional material

Most course material will be available via the GitHub organization hosting this document. Some material is hosted on the course's Canvas page, particularly material requiring a CU netID for access (e.g., course reserve books at the Math Library and online as eBooks, LinkedIn Learning video links).

The slides for all lectures will be made available via the LectureResources Git repo just prior to each class, as PDF files. Some lectures will involve work on a whiteboard; I'll make an effort to capture whiteboard content and included it in a post-lecture revision of the slides (accessible by simply doing a `git pull` in your copy of the LabResources repo).

We will not be following the content of any one book. Lecture and lab notes (mine and yours), as well as assigned readings from various sources, should suffice for completing the course with a high grade if you're a good note-taker.

That said, if you plan on using Bayesian methods in your career, you should invest in at least one good book on Bayesian methods. One that's close in spirit to this course is:

Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan
By John Kruschke

- Amazon.com
- Publisher's site (hardcover/eBook combo 50% off as of Dec 2021)
- Author's book site
- Author's blog
- Via BigWords.com (a new/used textbook search service)

Note it is the newer, 2nd edition, that I recommend. The library currently has only the first edition as a physical book; it will be available on reserve at the Math Library. However, both editions are available as eBooks.

Another introductory text, equally close in spirit, is:

Statistical Rethinking: A Bayesian Course with Examples in R and Stan
By Richard McElreath
(The book website also has example code for PyMC3, brms, & Julia)

- Statistical Rethinking – Richard McElreath's book site
- Amazon.com

The library has it available as an eBook (albeit in the awkward O'Reilly Online Learning Platform).

If you believe Bayesian methods will play an important role in your career, consider purchasing (or downloading) what has become the standard reference:

Bayesian Data Analysis, Third Edition
By Andrew Gelman, et al.

- Amazon.com
- Authors' site (with a PDF version, free for non-commercial use)
- Publisher's site

This book covers BDA at an advanced level, suitable for a statistics PhD student. However, much of it should be accessible to you, if not now, then at least after you've completed this course. Earlier editions may be available cheaply, but there is quite a bit of important material new to the third edition.

For Bayesian computation via *Markov chain Monte Carlo methods*, which will play a key role in this course, the following recent collection of tutorial chapters is becoming a standard reference:

Handbook of Markov Chain Monte Carlo
By Steve Brooks, et al.

- Amazon.com
- Authors' site
- Publisher's site

Many of the chapters cover advanced methods, beyond what we will cover, but there are also good chapters on the basics. The book is quite expensive, but *four chapters are available for free* at the authors' site; the first two are essential reading. This book will be on reserve at the Math Library.

Various scientific and engineering disciplines have produced books covering Bayesian methods tailored to specific fields. Cornell's library has quite a few of these; I'll mention some as the opportunity arises.

In my own fields of physics and astronomy, and in AI/computer science, one of the most influential texts on Bayesian foundations is:

Probability Theory: The Logic of Science
By Edwin T. Jaynes; ed. by G. Larry Bretthorst

- Amazon.com
- Publisher's site

Jaynes started working on this book in the late 1950s; I knew him and had access to some early versions. Unfortunately, he was such a perfectionist that he never felt happy with the book, and he left it unpublished at his death. His former student, Larry Bretthorst, did some final editing so the book could be published posthumously. The first three chapters are available on Bretthorst's web site: PTLOS chapters 1-3 (PDF).

This book offers probably the clearest and most thorough modern account of the principles of Bayesian inference, and fundamental analytical developments. It has little content relevant to computational implementation of Bayesian methods, but of course a deep understanding of foundations and fundamentals is a great help for practical use. The book is quite polemical in places (reflecting its history). Persi Diaconis, an influential mathematician and Bayesian statistician (and former Cornellian, and magician!), wrote a wonderful, frank, and very positive review that is worth reading:

"A Frequentist Does This, A Bayesian That" (Diaconis's review of Jaynes's PTLOS)

I've put several other useful books on reserve; see the Canvas site for a list.

Finally, I **strongly** recommend the following largely nontechnical book for those particularly interested in conceptual and philosophical issues arising in the foundations of statistics, particularly regarding how probability should be used to quantify uncertainty:

- Ten Great Ideas about Chance (Princeton University Press) by Persi Diaconis & Brian Skyrms
- Amazon.com: Ten Great Ideas about Chance (9780691174167)
- Review in *Notices of the AMS*: Ten Great Ideas about Chance — A Review by Mark Huber

This book grew out of a Stanford undergraduate course of the same name offered by Diaconis and Brian Skyrms, an influential philosopher of science. Although most of the text is nontechnical, many chapters have mathematical appendices, most of them quite accessible, but some a bit challenging.

Lecture plan

- Course introduction; models, measurements, arguments
- Foundations
  - Probability theory as logic
  - Key theorems: Bayes's theorem; the law of total probability
- Bayesian inference with discrete data
  - Binary hypotheses and data (Bernoulli distribution; binary classification)
  - Continuous parameter estimation with binomial data (binomial and beta distributions)
  - Parameter estimation with multinomial data (multinomial and Dirichlet distributions)
  - Parameter estimation with Poisson data (Poisson and gamma distributions)
- Bayesian inference with continuous data
  - Normal (Gaussian) and Student's t distributions
- Composite hypotheses
  - Propagating uncertainty
  - Model comparison & marginalizatino
  - Prediction & model checking
- Basic Bayesian computation
  - Laplace approximation
  - Quadrature and cubature rules
  - IID Monte Carlo integration
  - Markov chain Monte Carlo (MCMC)
- Multivariate relationships
  - Conditional dependence/independence, graphical models
  - Bivariate normal distribution
  - Regression
- Hierarchical Bayesian models
  - Shrinkage
  - Measurement error models
- Bayesian computation beyond the basics
- Assigning direct probabilities (sampling distributions and priors)
  - Consistency and symmetry requirements; functional equations
  - Jeffreys priors; reference priors
- Basics of decision theory & experimental design

The following is a *tentative* schedule with the early lecture topics:

Lec #	Date	Topic
1	Jan 25	Course intro; Motivation: Models, measurements, arguments
2	Jan 27	Probability theory as logic
3	Feb 1	Key theorems (Bayes's theorem; the law of total probability)
4	Feb 3	Bayesian inference with binary hypotheses and data
5	Feb 8	Continuous parameter estimation with binomial data
6	Feb 10	Parameter estimation with multinomial data
7	Feb 15	Parameter estimation with Poisson data
8	Feb 17	Parameter estimation with continuous data: normal distribution
9	Feb 22	Composite hypotheses: Model comparison & marginalization
10	Feb 24	Composite hypotheses, 2: Prediction & model checking

After the Feb break, we'll synthesize what we've learned to a general prescription for inference with parametric models, and then continue with more sophisticated models---Bayesian counterparts to multi-parameter conventional regression models.

Next we'll focus on Bayesian computation, culminating with Markov chain Monte Carlo (MCMC).

With flexible computational tools in hand, we'll explore richer model structures---*hierarchical Bayesian models* (also known as multilevel models, or probabilistic graphical models).

I have a menu of further topics we'll address as time allows. If you've encountered particular BDA topics you'd like to learn more about, please let me know; I may be able to work them into the schedule.

Lab plan

The labs will focus on the homework assignments, aiming to help you build the skills and insights needed to put the lecture content to work in solving problems.

For the first few weeks, the labs will operate somewhat separately from the lectures, aiming to build familiarity with the tools we'll use to implement nontrivial Bayesian computations later in the course.

Lab #	Date	Topic
1	Jan 28	Markdown, Git, GitHub
2	Feb 4	Jupyter notebooks
3	Feb 11	The PyData stack
4	Feb 18	Bayesian computation for single-parameter models

As the lectures move beyond fundamental, analytically tractable examples, the labs and lectures will mesh more strongly, with labs implementing computational methods and flexible models covered in lecture.