VANDERBILT
College of Arts *and* Science
Department of Physics & Astronomy

ASTR 3970 / 8070
Astrostatistics

Course description

The purpose of this course is to equip you with the knowledge and tools necessary to extract understanding and meaning from real datasets. Real-world data is not perfect; it is often not evenly spaced in time, there can be large gaps, sometimes the uncertainties have not been estimated correctly, there are often instrument calibration errors, and unknown processes that muddy the waters. Nevertheless, sophisticated statistical techniques exist to tackle such gremlins. We will study these techniques together, learning how we can cut the Gordian knot of complicated datasets to learn about pulsars, the cosmic microwave background, black holes, and more. The principles you will learn about are not all specific to astronomy/ astrophysics— they can in fact be ported over to many other disciplines and careers.

All of the statistics and data-mining that I know were gleaned from private study, and generally speaking I have found that this is true for many astrophysicists; we learn during our research careers for the task at hand. But only later do we realize that many techniques are related, and strategies used in one discipline can be brought to bear on others. It is therefore my intent to introduce you to the world of astrostatistics through the techniques that I have found most useful, and that you will likely use the most in your research. We won't dwell on the philosophy of inference; we're going to get our hands dirty.

Key topic blocks and learning outcomes

- The role of probability in inference
- Frequentist and Bayesian inference
- Bayesian parameter estimation and model selection using MCMC
- Exploratory data analysis and visualization
- Regression analysis
- Time-series analysis
- Machine learning and neural networks

Class logistics

- Tues, Thurs 1:15 PM – 2:30 PM
- In person, Stevenson 1 (Math), Room 214.
- Backup Zoom link available on Brightspace. *Will only be used when explicitly specified.*

Instructor

- Prof. Stephen R. Taylor, PhD
- Email: stephen.r.taylor@vanderbilt.edu
- Webpage: https://my.vanderbilt.edu/stephentaylor
- 17th & Horton Building, Office A1114B

ASTR 3970

- This class is cross listed with the graduate-level class ASTR 8070. All class sessions will be the same for both groups of students.

Office hours for homework feedback or help

- Wed *(online)*10:00AM—11:00AM Central Time.
- Fri *(in-person or online)* 3:00PM—4:00PM Central Time.
- NOTE: If you come to office hours to ask for homework help, please show that you have attempted the solutions beforehand.

Class participation and collaboration

- The first portion of class will have a lecture format, with me explaining the material. The second portion will be individual and collaborative problem-solving using Jupyter notebooks running on your personal machines.
- Participation credit will be assigned by submitting your completed copy of the lecture Jupyter notebook, with required tasks indicated therein. All completed lecture notebooks for the week must be **submitted by 11.59pm Central Time each Saturday**. Credit is given for making a reasonable attempt at all tasks in the notebooks.

Homework assignments

- Available to access on Fridays, and **due by 11.59pm the following Saturday.**
- However, HW 1 will be due at the end of **Wk 1 on Saturday, Jan 11th at 11.59pm**.
- We will use GitHub to submit classwork and homework. More details will be provided.
- There will be **8 homework assignments**.
- Homework will be graded along similar principles as Prof. Runnoe's ASTR 8020 class.
  - All solutions must include words to explain how the problem was solved. Solutions without adequate method commentary will have points deducted.
  - Code will be graded on (i) how well it is commented, (ii) how well it is structured, (iii) how well it is made compact and optimized, (iv) its speed, and (v) its efficacy in delivering the correct answer.

Final project

- You will complete a **capstone project** on a topic of your choosing. This should employ data-analysis principles learned in the course, and can involve data from your field of study. Note that this is not a semester-long project.
- You will write a **journal-style article** that includes an abstract, introduction, methods, figures, references, and appendix material. Your primary goal is to take something complicated and find a simpler way to explain it (in your own words). Don't just regurgitate information from the textbook or other resources.
- The length of the article should be 4–5 pages (not including references).
- The **appendix material should be in the form of a Jupyter notebook** that can reproduce the main calculations and figures of the article.
- The grading rubric for this project is as follows:
  - **By 11.59pm on March 17th**, e-mail me with your topic. **(10%)**
  - **By 11.59 pm on March 24th**, submit a 1-page abstract of your project to Brightspace. **(10%)**

- ○ **During class time on April 10th**, make a conference-style 10-minute presentation to the class on your project. (**20%**)
- ○ **By 11:59 pm on April 28th,** submit article PDF and appendix Jupyter notebook to Brightspace. (**60%**)**.** Assessed by the quality of background/motivation and explanations, appropriateness of figures/references, quality of writing and analysis.

Grading metric

- Class participation and collaboration (30%)
- Homework = 40%
- Final project = 30%
- Total = 100%
- A+ = more than 95%; A = 90-95%; A- = 85-90%; B+ = 80-85%; B = 75-80%; B- = 70-75%; C+ = 65-70%; C = 60-65%; C- = 55-60%; D = 50-55%; F = less than 50%

Textbook & Required Materials

- ***"Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data"***— *Ž. Ivezić, A. J. Connolly, J. T. VanderPlas, & A. Gray* (course textbook)

Other useful texts

- "Bayesian Logical Data Analysis for the Physical Sciences"— *P. C. Gregory*
- "Modern Statistical Methods For Astronomy"— *E. D. Feigelson & G. J. Babu*
- "Bayesian Data Analysis"— *A. Gelman, J. Carlin, H. Stern, D. Bunson, A. Vehtari, D. Rubin*
  - *Free:* http://www.stat.columbia.edu/~gelman/book
- "Practical Statistics for Astronomers"— *J. V. Wall & C. R. Jenkins*
- "Python Data Science Handbook"— *J. T. VanderPlas*
  - *Free:* https://github.com/jakevdp/PythonDataScienceHandbook
- "Information theory, inference, and learning algorithms"— *D. MacKay*
  - *Free:* https://www.inference.org.uk/mackay/itila/book.html
- "Data analysis recipes: Fitting a model to data"— *D. Hogg, J. Bovy, D. Lang*
  - https://arxiv.org/abs/1008.4686
- "Data analysis recipes: Probability calculus for inference"— *D. Hogg*
  - https://arxiv.org/abs/1205.4446

Policies

Class Attendance

- Real-time attendance is highly preferred, as it will allow for interactions and discussions that are essential to your understanding of this subject.
- If for some reason you are running late to connect to class, please still attend. Even if you are an hour late, please come. Coming to class late is better than not coming to class and you will not be judged. However, you will put yourself at a disadvantage if you do not have regular attendance.

- The lecture portion of classes will be a mixture of slides and chalkboard, with me explaining the material. You should make notes in real-time. The other portion of class will be individual and collaborative coding and problem solving.

Personal Issues

To ensure that concerns are properly addressed from the beginning, if you have a physical, learning, or psychological issue or disability and require accommodations, please let me know as soon as possible. You must register with, and provide documentation of your disability to Student Access Services.

There Will be No Extra Credit

Homework

- My grading philosophy is weighted in favor of you showing that you understand the problem(s). Be as explicit as you can in all your work, and show all steps in the calculation to receive full credit. **If you know your final answer is wrong or incomplete, say so!** This tells me that you understand the material.
- You are strongly encouraged to confer with your classmates on homework assignments, but I expect the work you submit to be your own. I will easily be able to use git tools to check whether code has been copied.
- All assignments (but *not their grades*) will be available to the rest of the class in the Git repository. My hope is that you will learn collaboratively from each other's approaches. Science is a collaborative endeavor: This statement is a fair warning that your work will be freely available to all other students in the class. Sharing knowledge and expertise in this way is intended in the spirit of scientific collaboration among your classmates. But to reiterate, I will easily be able to use git tools to check whether code has been copied.

Late submission of assignments or final project

Barring special arrangements made in advance of the due date, late submissions of portions of the final project will not be accepted for credit. Barring prior arrangements, late submissions for lecture notebooks and homework assignments will be subject to the following deductions: 1 day late −25%, 2 days late −50%, 3 days late −75%, 4+ days late will not be accepted for credit.

Generative AI

- Large language models like ChatGPT are readily available. However, they are not a replacement for your own thinking, reasoning, and comprehension.
- Use of ChatGPT, Gemini, Claude, or other AI chatbots to provide complete solutions for homework or the final project is prohibited.
- However, you may treat it like an expert colleague, and in fact, I recommend GitHub's Copilot for coding assistance. You may use it to help resolve misunderstandings, provide other examples or explanations, or other things that supplement the learning process. But remember, it can be wrong! If you use AI for something, then cite it appropriately.

Academic Honesty

- The Vanderbilt Honor Code applies to all graded work done in this class.

- I encourage you to freely discuss any or all content of the course with your peers, but **the work you submit must be your own**. Collaboration is encouraged, and you may show your broken code to a colleague and seek their advice, but students may not copy one another's homeworks or code. Any instance of academic dishonesty (including plagiarism) will be dealt with according to university regulations. Any form of academic dishonesty is unacceptable to our community and will not be tolerated. Students should report any suspected violation of proper academic behavior to me. I will report suspected violations of standards of academic honesty to my Department Head, and/or the Dean.

Course Outline

The following outline is meant as a guide only and subject to revision. The exact topics covered may deviate based on time constraints and class interest.

Section	Topics	Reading	Notes
Probability & Statistical Distributions	• Probability theory • Random variables • Probability and frequency • Central limit theorem • Generating random draws from arbitrary distributions	Ivezic Ch 1, 3	*Week 1-2*
Frequentist Inference	• Point estimation • Least squares estimation • Maximum likelihood estimation • Bootstrapping and jack-knifing • Comparison of distributions	Ivezic Ch 4	*Week 3-4*
Bayesian Inference	• Priors • Parameter uncertainty quantification • Model selection • Conditional distributions • Marginalization	Ivezic Ch 5	*Week 5-6*
Data Exploration & Visualization	• Non-parametric and parametric density estimation • Dimensionality reduction • Principal Component Analysis • Visualizing data	Ivezic Ch 6, 7	*Week 7-8*
Regression, parameter estimation, and model selection	• Formulating a model • Likelihoods • Markov chain Monte Carlo • Practical parameter estimation and model selection • Cross-validation techniques	Ivezic Ch 8	*Week 9-12*

Section	Topics	Reading	Notes
Time-series Analysis	• Deterministic and stochastic processes • Auto-correlation and cross-correlation • Structure function • Random Gaussian processes • Fourier power spectrum, Lomb-Scargle periodogram • Bayesian spectral estimation	Ivezic Ch 10	*Week 13*
Deep Learning	• Neural networks • Adding hidden layers • Fully connected, recurrent, and convolutional networks • Practical deep learning • Examples	Ivezic Ch 9	*Week 14*