VANDERBILT UNIVERSITY®
COLLEGE OF ARTS AND SCIENCE
Department of Physics & Astronomy

ASTR 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 2:20 PM – 3:35 PM
- Online via Zoom (link in Brightspace)
- Please ensure that you have adequate internet connection for audio and video participation.
- Lecture portions of class will be recorded for later review.
- **In-class reading days:** No new material will be presented on 02/23 or 04/08, but we will still meet.

Instructor

- Prof. Stephen R. Taylor, PhD
- Email: stephen.r.taylor@vanderbilt.edu
- Webpage: https://my.vanderbilt.edu/stephentaylor

- Tel: 615-343-6296 (office)
- Stevenson Center 6 (Physics & Astronomy), Office 6910 (9th floor)

Office hours for homework feedback or help

- Wed, Fri 10:00AM—11:30AM Central Time, or by appointment.
- Office hours will be conducted on Zoom. I will have an open Zoom line during these times (link in Brightspace), but it will help to let me know if you plan on connecting.
- 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 Friday**. 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 Friday.**
- However, HW 1 will be due at the end of **Wk 1 on Saturday, Jan 30th at 11.59pm**.
- We will use GitHub to submit classwork, homework, and the final take-home exam. Further details will be discussed on the first day of class.
- There will be **9 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 take-home exam

- After the last session of the class I will assign a take-home exam that includes a series of statistics, coding, data-mining, and data-analysis tasks that assess the course material.
- Exact date, time, and format TBD.

Grading metric

- Class participation and collaboration (30%)
- Homework = 40%
- Final take-home exam = 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 on the Zoom classes 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 the Zoom class, please still attend. Even if you are an hour late, connect. 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 virtual 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.
- The lecture portion will be recorded to review later. If you miss class or are viewing asynchronously, you must watch the relevant lecture video before the next class.

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.

Late submission of assignments or exam

Barring special arrangements made in advance of the due date, late submissions of the final exam 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.

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. It is your responsibility to avoid complaints or appearances of impropriety.
- Vanderbilt University is built upon a strong foundation of integrity, respect and trust. All members of the university community have a responsibility to be honest and the right to expect honesty from others. 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*
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*