Course preview

Trying to decide whether to take this course? In addition to the syllabus, you can have a look at the first lecture from the 2021 version of the course.

Teaching staff

Instructor: Vinothan N. (Vinny) Manoharan
email: vnm@seas.harvard.edu (but please use Slack rather than email for questions about homework or lectures!)
Office Hours: Fridays 2-3 pm on Zoom (https://harvard.zoom.us/my/vinny)

Teaching Fellow: Jennifer McGuire
email: jennifermcguire@g.harvard.edu
Office Hours: Mondays 1:00-2:00 pm Maxwell Dworkin 119

COVID Safety

We want to stay safe, so please follow these rules:

- **Wear a mask.** Per university policy, university-approved face coverings must be worn at all times in the classroom. Please make sure that your mask covers your nose and mouth at all times. If you cannot wear a mask, you must contact the Accessible Education Office (see the end of the syllabus) to make accommodations.
- **No eating/drinking.** Per university policy, there is no eating or drinking in the classroom. If you need to take a drink, please feel free to leave the classroom at any time and return.
- **No visitors.** Only registered students are allowed to attend class. No visitors are allowed. If you are interested in auditing the class, please contact the instructor first for permission and guidelines.

Course meeting schedule

Class will meet in Lyman 425, 10:30-11:45 am MWF. I will usually lecture on Mondays and Wednesdays. Fridays are reserved for section, with some potential exceptions for make-up lectures (these will be announced).

Course aims

This is a course about data in physics experiments and how to draw conclusions from it. Most physics courses start from general physical laws (for example, Maxwell's equations) and derive specific predictions from them. That process is called *deductive inference*. But as a PhD student you are expected to contribute to the discovery of new laws and concepts. This course aims to teach you the techniques for reasoning from the data to determine the validity of a particular theory or model, or to determine the most likely value of a parameter (for example, the percentage of dark matter in the universe) for a given model. This process is called *statistical inference*. It is fundamentally different from deductive inference but just as important, and all experimentalists need to be familiar with it.

Doing statistical inference on modern data sets requires a computer and tools more powerful than a spreadsheet. This course therefore covers not only statistical methods, but also the methods of dealing with data on the computerâ€”including loading, filtering, plotting, visualizing, and simulating it. We'll do everything in Python, because it is a general-purpose language, it is easy to learn, and it has powerful tools for data analysis. It's also free.

This course assumes little about your ability to program. We will start from the very basics and build up to advanced calculations.

Learning objectives

The main objective is to prepare you for research. By the end of the course, you should be both competent and confident in using the tools of statistical inference to analyze experimental results and derive conclusions from them. You should also be able to critically analyze published results. We will focus on Bayesian approaches.

Modern statistical inference relies heavily on computation. By the end of the course, you should be able to program proficiently in Python and follow good programming practices, including vector-based computation, modular code, and revision control. Through the final project, you'll become familiar with tools for collaborating on code, and you'll learn how to write well-documented code that can be easily

shared with others.

Another objective is for you to become familiar with the types of data and data analysis used in other subfields. To this end, many of our classes will include discussions, so that you can learn from your classmates. Participation is therefore essential to your learning in this course.

Is this course for you?

If you are an experimental physicist, physical scientist, or engineer who is familiar with the process of obtaining experimental dataâ€”including designing experiments to minimize systematic error, doing experiments, and estimating uncertaintiesâ€”then yes, this course is for you.

If, however, you do not have any background in doing experiments, then the course is probably not for you. I would argue that one should first learn how to do experiments before learning how to analyze the data from them. All the fancy data analysis methods in the world won't help you if you cannot critically evaluate the methods used to obtain the data.

More specific advice: If you are

- *An experimentalist with good background in numerical and computational techniques*: You might find the early part of the course slow-going, in which case you might prefer to take a course such as ENG-SCI 255 or APMTH 207. Both of these courses deal with statistical inference, though in different contexts (not necessarily physics). Both also assume that students come in with experience with programming. Another course with a statistical inference component (in a biological context) is MCB198/AM215.
- *A theorist:* If you already have significant experimental experience, you'll find it useful. Otherwise I would recommend that you take a laboratory course such as PHYSICS 191R or PHYSICS 247R first.
- *An undergraduate*: as above, if you have experimental experience (in a research context), then yes. If you are not doing research, then no.
- *Interested in data science:* Our approach is different from that of data science, in that we are generally testing mechanistic models or theories. Students interested in data science might want to take APCOMP 209.

Outline of topics

List subject to change:

- Introduction to Bayesian and frequentist inference
- Bayes' theorem and how to apply it
- Bayesian parameter estimation and hypothesis testing
- The maximum-entropy approach
- Linear and nonlinear model fitting
- Markov-chain Monte Carlo methods

If time permits, we *might* also discuss the following:

- Frequentist hypothesis testing (including discussion of *P*-values and replication crisis)
- Causal inference
- Time-series analysis
- Hierarchical Bayesian models
- Machine learning and physics

Textbook

There are two textbooks. Both should be available at the COOP and can be purchased using this link. Both can also be purchased as eBooks:

1. Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support, by Phil Gregory (Cambridge University Press). See also errata for the paperback edition and errata for the original printing. Note also that you can get the eBook version through the Harvard Library (http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1017/CBO9780511791277).
2. A Student's Guide To Python for Physical Modeling, by Jesse M. Kinder and Philip Nelson (Princeton University Press). **Note: I recommend the 2021 edition, and not the 2018 or 2015 editions**. See also the book webpage for errata, examples, and updates. This book is optional, but if you are new to Python I strongly recommend it.

Other sources, which will be placed on reserve at Cabot library, include

- *Statistics for Nuclear and Particle Physicists*, by Louis Lyons (Cambridge University Press)
- *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, by R.J. Barlow (Wiley)
- *Statistical Data Analysis*, by Glen Cowan (Oxford University Press)
- *Data Analysis: A Bayesian Tutorial*, by D.S. Sivia with J. Skilling (Oxford University Press)
- *Causality: Models, Reasoning, and Inference*, by Judea Pearl (2nd ed, Cambridge University Press); available online through the library

Assignments

Homeworks: Homeworks are assigned weekly. These assignments will involve coding and inference. There will also be some short assignments that consist of brief presentations or peer reviews of code. Extensions on homeworks are at the discretion of the TFs, but please see the "Isolation/Quarantine Accommodations" section below.

Participation: Participation and discussion are essential to this course. Participation scores are based on presentations, peer reviews (which are done outside of class), and questions or other contributions during lecture. Students will not be penalized for missing class due to isolation or quarantine.

Project: During the second half of the course, you will do a final project involving the analysis of actual data (either obtained by you or available elsewhere). Ideally, the project is ambitious enough that it could *eventually* lead to a publication, but not so ambitious that it will take you more than a month to do it. The project will be structured so that you will get feedback at each step.

Grading

You will submit homeworks online as Jupyter notebook files. Only notebooks in which all cells execute without errors can be graded. We recommend running "Restart Kernel and Run All Cells" to check for errors before you submit. For full credit, your notebook should be well documented, and the intent of the code should be made clear.

Homeworks will count toward approximately 40% of your grade, participation 20%, and the final project 40% (values subject to change).

Section

Whereas the lecture component will give background and information on the analysis techniques, section will cover implementation. We will focus on good programming practices and learning to use the most recent Python tools.

Collaboration policy

See the course's academic integrity policy. Also please review the GSAS Handbook (https://gsas.harvard.edu/codes-conduct/academic-integrity) for general information on academic integrity.

Getting help

The teaching staff is here to help! We have office hours, and we monitor the Slack workspace for questions. Please post questions about lecture material or homework problems in one of the public Slack channels (we will set up a separate channel for each homework) rather than direct messaging, since other students may have the same question.

We also have a wiki and other resources for the course on GitHub.

Quarantine/Isolation accommodations

Given the high COVID caseloads and positive-test rates, we recognize that many students may have to quarantine or isolate. Furthermore, if you are feeling unwell, please do not come to class. There is no penalty for missing class due to symptoms, quarantine, or isolation, though you should inform the instructors if you are unable to come to class so that we can make accommodations as follows:

For students in isolation or quarantine, we will set up a Zoom stream and/or recording of the class. The quality may not be ideal, so we also encourage all students to identify a "study buddy" at the start of the semester. Your study buddy can provide you with lecture notes. I will also post official lecture notes but it usually takes me 5-7 days to prepare and upload them.

We will also grant extensions on assignments for students in isolation/quarantine. Please be in touch with the TF about how much extra time you will need.

Accommodations for students with disabilities

Students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with the professor by the end of the second week of the term. Failure to do so may result in the Course Head's inability to respond in a timely manner. All discussions will remain confidential, although Faculty are invited to contact AEO to discuss appropriate implementation.