STAT 102: Introduction to Statistics for Life Sciences

Spring 2024

Meeting Time: Tues/Thurs 10:30 - 11:45 AM
Course Site: https://canvas.harvard.edu/courses/127545

Instructor: James Xenakis jxenakis@g.harvard.edu
Preceptor: Julie Vu julievu@g.harvard.edu

Course Description

Statistics has become an integral part of research in medicine and biology, and the tools for summarizing data and drawing inferences from data are essential both for understanding the outcomes of studies and for incorporating measures of uncertainty into that understanding. This course motivates statistical methods through data analysis and visualization, in addition to examining the underlying theory. The topics discussed include study design, exploratory data analysis, probability, inference, modeling, and data ethics.

As a course in applied statistics, this course aims to prepare students to be responsible practitioners of the discipline, with sufficient technical training to learn and apply field-specific methods as well as a strong sense of statistical literacy and awareness of how statistics can be misused. This course uses the statistical programming language R; students will gain fluency with using R for conducting data analysis.

This course is designed for students who are interested in working in medicine, public health, and the life sciences. The majority of examples and exercises in the course are based around published data associated with peer-reviewed research in these fields.

Prerequisites

There are no prerequisites. Students with prior exposure to introductory statistics (such as through an AP Statistics course) will encounter some overlapping material but gain a stronger understanding of the statistical thinking and principles behind commonly used methods, in addition to learning practical skills for analyzing data. No prior experience with a programming language is required; the first few weeks of the course have a strong focus on acquiring R skills.

1 Learning Objectives

The primary goal of this course is prepare students to conduct research in an applied field, which requires being comfortable with 1) the technical details of analysis, 2) statistical computing, 3) and the scientific process as a whole, from determining appropriate methods for addressing a research question to clearly communicating findings in the context of the question. Furthermore, this course will provide students with the statistical literacy skills to critically interpret research findings from studies and media reports.

1. *Technical skills.* This course is an introduction to statistical inference, which can be broadly defined as methods for drawing conclusions about features of a population based on data from a random sample drawn from that population. In this course, we discuss inference in the context of numerical outcomes, categorical outcomes, and regression modeling.

2. *Computing.* Computing is an essential part of the practice of statistics in the modern era. Even the simplest arithmetic calculations such as taking the mean of a set of numbers can become time-consuming without the use of computing. Exploring data, creating visualizations, and conducting analyses requires some level of fluency with a statistical computing language such as R.

3. *Scientific process and communication.* The most challenging part of investigating a scientific question usually lies not in the details of executing a particular analysis method, but rather with decisions that take place before and after data analysis. For example, which statistical methods are appropriate for analyzing a particular set of data? After conducting a set of analyses, how can the quantitative results be translated into a cohesive conclusion?

After taking this course, many students immediately proceed to applying these skills in a research setting. While many domain-specific techniques are beyond the scope of an introductory course, our intent is that this course provides students a solid foundation that makes it easier to acquire more advanced techniques, whether that might be coding simulations to model evolutionary dynamics or learning methods for analyzing time-to-event data.

2 Learning Environment

2.1 Class Sessions

Class sessions are conducted according to a 'semi-flipped' format. There is recommended pre-reading for each class session; while the pre-reading is not mandatory, we believe that completing the pre-reading will foster more effective learning during class. Each class session consists of a brief (30-35 minute) lecture, with the remaining time devoted to interactive labs. During the lab portion of class, students will work on problems as a group with classmates (with the teaching staff available to provide support), then regroup as a class to discuss solutions.

This course format is designed to foster a collaborative learning environment. We expect all students to attend class and participate. We understand that being asked to work on problems during

class can feel intimidating, especially compared to simply listening to a traditional lecture. However, we believe grappling with challenging material is ultimately the way that learning happens. Additionally, there is some evidence that students in classes with active learning perform better on exams than students in classes with traditional lecturing[1] and that active learning can decrease the achievement gap for underrepresented minorities and first generation college students.[2]

The class labs are not submitted for a grade and the solutions are posted immediately after class.

2.2 Sections

Weekly sections offer an opportunity for students to review concepts and work on practice problems in a smaller setting than class sessions. Students will be assigned to a 60-minute section.

Section attendance is optional, but highly recommended. The solutions to section problems are posted after sections take place.

2.3 Office Hours

The teaching team will host regular drop-in office hours each week. During office hours, feel free to ask conceptual questions or specific questions about problem sets, chat about statistics as a discipline, etc.

1-on-1 office hour appointments (15 minutes per session) are also available. These are useful for asking about a concept in more detail or discussing course progress. Please email to schedule an appointment if all available times conflict with your availability.

2.4 Discussion Forum

To foster a sense of class community, we will use Slack as an enhanced discussion forum. We encourage using Slack to communicate with fellow students and the teaching team, such as by asking questions about material covered during class or about the course assignments. In order for Slack to be a useful resource, please consider posting questions on public channels (rather than privately messaging the teaching team) so that other students can view questions and answers; this will also allow other students to contribute answers.

The course staff will moderate Slack on a daily basis. Please be mindful of the honor code while using Slack—for example, avoid sharing written answers word-for-word. Detailed instructions for using Slack will be provided on the course site, along with our community guidelines.

[1] Freeman, et al (2014)
[2] Theobald et al (2020)

3 Course Materials

3.1 Readings

Readings will primarily be assigned from *Introductory Statistics for the Life and Biomedical Sciences*, by Vu and Harrington. There will also be assigned readings from *ModernDive*, by Ismay and Kim, and *Introduction to Modern Statistics* by Çetinkaya-Rundel and Hardin. All readings are free, open-source texts and links are posted on the course site.

In addition to textbook readings, students will also be introduced to news articles, published literature, journal editorials, etc. on relevant statistical topics.

3.2 Computing

The course will use the statistical language R via R Studio, which is freely accessible with the cloud-based interface Posit Cloud (https://posit.cloud/). This allows all computing to be done within a web browser with internet access.

R and *RStudio* are also freely available for all common operating systems and instructions for downloading R, *RStudio*, and LaTeX are available on the course site.

4 Assessments

The course grade will be based on problem sets, a midterm exam, a final exam, and participation, with the following weights:

Component	Weight
Problem Sets	30 %
Participation	10/15 %
Midterm	25 %
Final Exam	35/30 %

The distinction between the 10-35 and 15-30 weighting for participation and the final exam is explained below.

4.1 Problem Sets

Problem sets will be submitted electronically to Gradescope. Graded problem sets with commentary will be available before the next problem set is due. There will typically be a problem set due each week on Thursday, at 11:59pm EST; refer to the course calendar below for the specific deadlines. Problem sets are posted at least seven days prior to the due date. Solutions to the problem sets will not be posted; be sure to review the individualized feedback and ask clarification questions about any incorrect answers.

In order to be accommodating of everyone's personal situation, we will adhere to the following polices:

- The lowest problem set score will be dropped from the grading. This includes scores of 0, such as for a problem set that was not submitted.
- Each student has four extension days that can be used as needed, no questions asked. Using 1 extension day means that the submission deadline is extended by 24 hours. To use an extension day, notify your assigned section leader. We recommend that you use no more than one extension day on a problem set in order to avoid falling behind, but recognize that there may be circumstances in which using more than one extension day may be necessary.
- Once the extension days are used, no further extensions will be granted. In the event of serious illness or unexpected family circumstances that may require additional flexibility, please contact James and Julie and provide a note from your resident dean as documentation.

Regrade Requests: Problem set regrade requests must be made on Gradescope within a week of receiving the score and grading feedback. In the request, be sure to clearly and succinctly state what error you believe occurred. The whole problem, and possibly the whole assignment, will be re-graded. Your grade may increase, stay the same, or decrease.

4.2 Exams

Both the midterm and final exams will include an in-class component (3 hours) and an oral component (10 minutes). Please bring a laptop to the in-class components. The oral component will take place over Zoom.

Students will not be allowed to collaborate on exams, and doing so will be treated as a violation of the Honor Code. Both exams will be open-book, open-notes. The in-class components will involve the use of computing to analyze datasets.

- The midterm exam should only be missed due to extenuating circumstances. There will be no makeup midterm exam; the final exam will comprise 60% (or 55%) of the course grade for students who cannot complete the midterm. Contact James and Julie if you anticipate needing to miss the midterm exam.
- The final exam will take place according to the College Final Exam Schedule.

4.3 Participation

The participation component of the course grade will be assessed in the following ways:

1. Structured discussion on Slack. Each problem set will include a discussion prompt, to be answered in a channel on Slack. To receive full participation credit, write a post in the Slack channel and respond to someone else's post before the problem set deadline. Each discussion prompt assignment is worth 5 points and scored on thoughtfulness. Note that discussion posts should still be submitted before a problem set deadline even if the problem set is submitted late.

2. Engagement on Slack: Post at least two messages on Slack before March 05 and post at least two messages on Slack before April 25. Messages that count include asking a question about course content, answering someone else's question, posting a useful resource and why you found it helpful, and creating an example that illustrates a recent concept. Note that the Slack prompt-based discussion posts do not count toward this engagement requirement.

3. Engagement at Office Hours: Attend at least one office hour before March 05 and attend at least one office hour before April 25. Students are not required to stay for the entire duration of the office hour to receive engagement credit. Come introduce yourself, ask a question about course content, talk about statistics outside of Stat 102, etc.

4. Attendance: You are expected to attend class. While class attendance is not explicitly factored into your participation grade, attendance is still strongly recommended. Section attendance is optional, but we do recommend that you regularly attend section if your schedule allows. If you have a schedule conflict, we suggest making time to review the section materials and bringing questions to office hours. Note that in past semesters, students have found section to be a very helpful component of the course and have even stated that they wished section were mandatory (so that they would be more motivated to attend).

5. Optional "Statistics in the News" group commentary. In self-chosen groups of 2-3, find an example of statistics in the news that relates to the ideas discussed in class and collaboratively write a brief commentary to be posted on Slack. This is meant to be a non-stressful activity that helps you meet other students in the course. The commentaries are due on April 25 and may be submitted at any time during the semester.

For students who opt out of doing a group commentary, participation is weighted 10% and the final exam is weighted 35%.

For students who choose to do a group commentary, participation is weighted 15% and the final exam is weighted 30%.

5 Course Climate

All members of the class will agree to abide by the following community norms:

> We pledge to make participation in our community a harassment-free experience for everyone, regardless of age, body size, visible or invisible disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, religion, or sexual identity and orientation.
>
> We pledge to act and interact in ways that contribute to an open, welcoming, diverse, inclusive, and healthy community.
>
> *This Code of Conduct is adapted from the Contributor Covenant, version 2.0.*

6 Academic Integrity

We encourage you to collaborate with classmates while working on problem set questions, but you must be sure to understand a solution well enough to be able to work a similar problem on your own. Solutions must always be written in your own words; this also applies to any program code. Copying or paraphrasing someone else's solution is a violation of the Harvard Academic Integrity policy. You are allowed to use R functions that are not covered in this course, but if your code deviates significantly from that taught in Stat 102, we may schedule a meeting to discuss your work and ensure that the work is your own.

Solutions to problem sets from previous versions of the course may be available in various places online. Copying answers from solution sets online is a violation of the Honor Code and any instances of doing so that we detect will be reported to the Honor Council.
The Harvard College Honor Code states:

> Members of the Harvard College community commit themselves to producing academic work of integrity – that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.

More information about the Honor Code as well as resources for students may be found at https://honor.fas.harvard.edu.

Students who sell, post, publish, or distribute course materials without written permission, whether for the purpose of soliciting answers or otherwise, may be subject to disciplinary action, up to and including requirement to withdraw from Harvard College.

7 Generative AI Policy

While generative artificial intelligence (AI) tools such as ChatGPT are capable of generating code, analyzing data, and producing written summaries, this course is intended to help students develop their **own** abilities to write code, analyze data, and thoughtfully communicate results. Therefore, we expect that all work (including code, written work, oral assessments) that students submit will be their own. We specifically forbid the use of generative AI tools to answer assessment questions, unless the assignment specifically states that it is allowed. Violations of this policy will be considered academic misconduct.

The purpose of this policy is not to lessen student access for support but to ensure that students gain important skills. While we recognize that AI tools can be powerful assistants, existing software has been shown to be error-prone in many cases, including misstating facts and even

completely fabricating research results. We believe that responsibly and effectively using AI tools requires some base knowledge (such as applied domain knowledge or coding experience) in addition to informed skepticism and critical thinking. Students are welcome to explore the use of AI tools in the learning process, such as using AI to find research papers related to course topics. Please contact James and Julie with any questions about the use of generative AI in this course.

Note that different classes at Harvard may implement different AI policies and students are responsible for conforming to course-specific expectations.

8 Accessibility

Harvard College is committed to working with all students. If you have a disability and would like to request accommodation for this reason, please contact the Accessible Education Office (https://aeo.fas.harvard.edu/). Advance notice and appropriate documentation are required for any accommodations.

9 Course Schedule

Date	Unit		Class	
01/23 (T)	0		Course Logistics	
01/25 (Th)	1	Introduction to Data	Data Collection Principles	**Pset 0 Due**
01/30 (T)			Summarizing Data	
02/01 (Th)			Summarizing Data	
02/06 (T)			DDS Case Study	
02/08 (Th)	2	Probability	Intro to Probability	**Pset 1 Due**
02/13 (T)			Conditional Probability	
02/15 (Th)	3	Distributions of Random Variables	Discrete Random Variables and the Binomial	**Pset 2 Due**
02/20 (T)			Normal Distribution	
02/22 (Th)	4	Foundations for Inference	Sampling Variability and Confidence Intervals	**Pset 3 Due**
02/27 (T)			Hypothesis Testing	
02/29 (Th)	5	Introduction to the Tidyverse	Data Visualization	**Pset 4 Due**
03/05 (T)			No Class (Midterm Exam)	
03/07 (Th)			Data Wrangling	
03/19 (T)	6	Inference for Numerical Data	Comparing Two or More Means	
03/21 (Th)			Power	**Pset 5 Due**
03/26 (T)			P-Value Pitfalls	
03/28 (Th)	7	Simple Linear Regression	Intro to Regression	**Pset 6 Due**
04/02 (T)			Inference in Regression	
04/04 (Th)	8	Multiple Linear Regression	Intro to Multiple Regression	**Pset 7 Due**
04/09 (T)			Inference and Interaction	
04/11 (Th)	9	Inference for Categorical Data	Inference for Proportions	**Pset 8 Due**
04/16 (T)			Inference for Two-Way Tables	
04/18 (Th)	10	Logistic Regression	Logistic Regression	
04/23 (T)		Data Ethics	Data Ethics and Course Wrap-Up	
04/25 (Th)		No Class (Reading Period)		**Pset 9 Due**