Harvard University, Statistics 117

Data Analysis in Modern Biostatistics

Syllabus, Spring 2023

Lectures: Tue/Thu, 9:00am–10:15am in Science Center 705
Instructor: Giovanni Parmigiani
E-mail: Please use canvas email
Course web site: `https://canvas.harvard.edu/courses/129153`
Office Hours: Thu 10:30am-11:45am in SC 316.06, or by appointment
Section Time:
TF: Ahmad Abdel-Azim
TF Office Hours: To be determined based on a student poll in the first week.

Overview: The course is an introduction to the application of statistical concepts in biomedical research, through the lens of biomarker research in cancer biology and medicine. The foundation are two comprehensive collections of data on gene expression and clinical characteristics for patients with cancer of the ovaries [*CuratedOvarianData* `LINK`] and breast [*CuratedBreastData* `LINK`]. The notes for Lecture 1 on the Canvas site provide an overview of goals and methods covered.

Objectives: To develop a sense for practical applications of statistical thinking and tools in biomedical sciences. To learn how to make and defend statistical modeling choices in complex realistic applications. To constructively critique the analyses of others. To develop a working knowledge of R tools relevant for biomedical data science applications.

Prerequisites:
Required: Stat 110 and (AP Stat or 102 or 104)/111/139.
Recommended: Stat 115; Basics of R programming.

Lectures and Sections: The course uses a combination of lectures, required sections, online discussions, individual homeworks, and both individual and group projects. Emphasis is placed on peer-to-peer discussion / critique and class presentations.

Lecture notes, including discussion topics will be shared at least one week ahead of the class. Students are expected to review the notes and post comments and questions on Canvas prior to class by noon the day before class. The instructor will review these comments and use them to lead discussion in class. Clarifications will be abundant, but lecture will be hard to follow without having read the notes.

Lectures will focus on discussions and project critiques. Sections will cover hands-on examples and programming. Attendance is mandatory for both.

One-hour weekly sections will begin the second week of the course. Sections will be the place to work through examples, review lecture material in a more detailed way, solve problems, and learn how to use the statistics packages for implementing the various methods taught in the course.

Homework: The workload will consist of four homework assignments and three projects. Homework will help you master analysis and coding skills to carry out the projects.

Homework assignments will be posted on the course's canvas page. You will be told when the assignment is posted online. Homework solutions are due by midnight (EST) on the due date. You are free to discuss and work on homework problems with other students, but you should write up your solutions independently (see the collaboration policy statement).

The official course policy is that no late homework will be accepted. In return for your timely submission of homework, we will make every effort to return graded homework promptly.

Homework Grading. Homework assignments will be graded by a grader or TF on a scale from 1 to 5. Homeworks are graded in large part on the clarity of your presentation of the solutions, not just their correctness. Homeworks that are generally clear and correct will earn scores of 4 or 5; those less so will earn a 3. Sloppy, incorrect and/or incomplete homeworks will receive a 1 or 2. All homeworks will count toward your course grade – we will not drop any homework grades.

Project Assessment and Grading. The course requires three projects. Projects will be graded by the TF on the same scale as the homework.

Separately, projects will also be reviewed by a peer group, and some will be discussed in class. After projects are turned in, we will form mini-groups (4-5 students) to discuss each other's projects. Each group will hold a guided discussion of project results, and nominates one project for class discussion in the next lecture. The instructor will review nominated projects and choose two or three. No additional credit will be given for this selection as they will be identified for controversy/discussion value as much as quality.

The final project will be due sometimes after the end of classes. Students will make a class presentation of the final project's results. We will set aside three lecture slots at the end of the semester for this. The final project will incorporate any feedback, and a discussion of how it was addressed. Additional more specific instructions will be given when the final project is posted.

Exam: The course will have one take-home midterm exam during the semester. The midterm will test methodological concepts from the first half of the course. Students will be required to take it online in a three-hour session within a 24-hour window.

Class Participation: Class participation is important and will be evaluated based on contributions to the canvas discussions, team project presentations, and class discussions.

Grades: Course grades will be determined by the following components, with the weights shown:

Homework assignments	20%	(5% each)
Midterm Exam	20%	
Course Projects	20%	(10% Each)
Class Participation	10%	
Final Project including Presentation	30%	

Computing: The computer package for this course is R, which runs on Windows, MacOS, and Linux systems, and can be used natively or within the Rstudio environment. R is free and available online through `www.r-project.org`. The free version of Rstudio is at `https://www.rstudio.com/products/rstudio/download/` R is straightforward to learn, but is sufficiently powerful and versatile to be useful for many projects that you might carry out after this course. Reference guides on R and on the specific R packages required by the projects will be placed on the course web site. Only basic prior knowledge of R is needed. General concepts and examples will be provided, but a fair amount of independence is expected in figuring out details. The online resources are very helpful.

Student are expected to turn in homework and project solutions in an R markdown language. RStudio offers both traditional R markdown and Quarto —which is the language used to develop the class materials.

Quarto supports python and you may use python functionality for your final project.

Course Materials There is no textbook for the course. Lectures will often include additional reading. Students are expected to fill most gaps on their own, but are always welcome to discuss further readings on canvas or in class.

Helpful materials include general references on computing and modeling, such as,

W. N. Venables, D. M. Smith and the R Core Team. *An introduction to R.* `free download`

G. James, D. Witten, T. Hastie, R Tibshirani. *An Introduction to Statistical Learning, with Applications in R.* `free download`

G. Parmigiani. *Modeling in Medical Decision Making.* Wiley 2002.

as well as articles or book chapters specific to the individual topics. All course materials will be free and available through the course website.

Canvas. All course documents, including homework assignments, supplementary material, links to data and software, will be available on the course web site on canvas. Students will receive course announcements through canvas e-mail. Homework and projects will be submitted through canvas.

Disability accommodations for timed exams. 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 instructor by the end of the third week of the term. Failure to do so may result in us being unable to respond in a timely manner. All discussions will remain confidential.

Collaboration policy statement. University policies against plagiarism will be strictly enforced. You are encouraged to discuss problem sets with your classmates, but each student must write up and code solutions separately. Be sure that you have worked through each problem yourself and that the answers you submit are the results of your own efforts. You also may not share or view another student's computer code, submit output from another student's computer session, or allow another student to view your code or output. A good rule of thumb: if a fellow student asks if you would like to discuss a homework problem, we encourage you to say "yes"; if a fellow student asks to see your answer to a homework problem or R code, the answer is "no."

You use it, you cite it. If you use AI like chatGPT to generate text or code for canvas discussions, projects or class presentations, you need to acknowledge the source and put the code or comment in quotes (or equivalent) the way you would if you were citing John Tukey or Aristoteles.

Using chatGPT text or code w/out acknowledgment will be considered cheating.