Stat 343: Mathematical Statistics

About the Course

Basic Information

- Meeting Time and Location: Mon, Wed, and Fri 11:00AM - 12:15PM in Clapp 401.
- Course Website: `http://www.evanlray.com/stat343_s2018/`
- Email: `eray@mtholyoke.edu`
- Office: Clapp 404C
- Office Hours: I will hold regularly scheduled office hours each week at times to be selected by you. These times will be posted on the course web site. Please do not hesitate to contact me to set up appointments for additional office hours at any time!

Textbook We will be using "Mathematical Statistics and Data Analysis" (3rd edition, ISBN 978-8131519547) by Rice as the primary text for this class. A copy will be on reserve, and the department has a copy you can borrow for short times. I will assume that you are familiar with most of the material in the first 5 chapters of this text, but we will review some essential topics in the first few days. We will really begin with Chapter 7, and occasionally review material from earlier chapters as necessary. This is a good book, but it doesn't cover all the material I would like to. In order to fill in the gaps in Rice, I will also be drawing on material from other sources. You don't need to purchase these other texts, but in case you are curious – additional readings will likely come from the following texts, among others:

- Introduction to Statistical Thought by Lavine
- Mathematical Statistics with Resampling and R by Chihara and Hesterberg

Piazza We have a Piazza page for this course at `https://piazza.com/mtholyoke/spring2019/stat343`. I ask that you **please submit all questions about course content as questions on the Q&A forum at Piazza**. This will allow other students to answer your questions if they see them before me, and will allow other students to benefit from the answers to your questions. Additionally, **you can post questions and answers to Piazza anonymously**.

Description The purpose of this class is to provide you with the theoretical understanding and computational skills necessary to select and implement appropriate methods for statistical inference in new settings.

We will examine the theory behind statistical inference procedures including point and interval estimation and hypothesis testing, as well as the computational methods needed to implement these inference procedures. In terms of theory, our goals are to understand common methods for deriving inference procedures for statistical model parameters, how and why these methods work,

and how we can evaluate the relative performance of different inference procedures. We will focus primarily on inference for parametric models, but will also discuss inferential techniques that can be used when parametric assumptions are suspect. Computational tasks will include implementation of estimation procedures such as Newton's method for maximum likelihood, MCMC methods for Bayesian inference, and the non-parametric bootstrap. We will use simulation studies extensively to compare the performance of different approaches to inference.

A tentative schedule and topic list is below. This outline is ambitious; we may not actually get to everything outlined here. An up-to-date list of topics covered so far and a time line for upcoming classes will be kept on the course website.

Unit	Topic	Weeks
Review and Overview	**Review**: Topics from probability, R	1, 2
	Survey Sampling: We will develop many of the central ideas of the course in the context of simple random samples from a finite population: sampling distributions, bias, variance, mean squared error, and a first look at confidence intervals.	1, 2
Point Estimation	**Frequentist Methods**: Method of moments; maximum likelihood via analytic maximization; Maximum likelihood estimates via numeric optimization; sampling distributions; bias, variance, mean squared error	2, 3, 4
	Bayesian Methods: Prior and posterior distributions; conjugacy; MCMC	4, 5
Large-sample Results	**Frequentist Results**: The asymptotic distribution of the MLE; Fisher information; Efficiency; the Cramér-Rao lower bound	6, 7
	Bayesian Results: Laplace approximation to the posterior distribution; large-sample convergence	7
Spring Break	**Have a good break!**	8
Interval Estimation	**Bayesian credible intervals**: Posterior percentiles; highest posterior density	9
	Frequentist confidence intervals: Exact; from large-sample approximation	9, 10
	Bootstrap confidence intervals: Non-parametric bootstrap confidence intervals; bootstrap t confidence intervals	10, 11
Hypothesis Testing	**Frequentist Tests**: The frequentist set-up; p-values, errors, power, and power functions; likelihood ratio tests; t and F tests	12, 13
	Connections between hypothesis tests and confidence intervals	13
Extra Week	If time, we will discuss special topics to be selected from mixture models, kernel density estimation, and shrinkage estimators	14, 15

Policies

Attendance Your attendance in class is crucial, unless you are sick. If you are sick, please let me know and stay home and rest; I hope you feel better!

Collaboration Much of this course will operate on a collaborative basis, and you are expected and encouraged to work together with a partner or in small groups to study, complete homework assignments, and prepare for exams. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or large blocks of R code from another student is not acceptable and will receive no credit or a penalty. No interaction with anyone but the instructor is allowed on any exams or quizzes. All students, staff and faculty are bound by the Mount Holyoke College Honor Code.

To sum up: **I want you to work together** on homeworks and labs. *But,* **you must write up your answers yourself.**

Cases of dishonesty, plagiarism, etc., will be reported.

Technology

Computing with R Modern statistics can't be done without computation. We will use the R statistical programming language in this course. R is one of the most commonly used programming languages in academic statistics, and I use it daily; it's also very commonly used in statistics and data science positions in industry. Knowing R is a marketable skill. In this class, you will use R most days, and for many homework problems. I expect that you are familiar with R from previous classes, but I do not expect that you are an expert at R yet. That said, it is imperative that you let me know if you are confused about anything we are doing in R as soon as possible.

We will use R via RStudio; Mount Holyoke's version of RStudio Server can be accessed at `https://rstudio.mtholyoke.edu/`. You are also welcome to work locally on your own computer if you have RStudio set up; however, please make sure you have installed at least version 3.5.0 of R and the latest versions of any R libraries we use.

It will be important to **bring your laptop to class**; we will be using R nearly every day. Much of this work will be done in pairs, but we need to ensure that there is a sufficient number of computers. Please let me know if this presents any issues, as there are laptops available for you to borrow.

Version Control with Git and GitHub Git is a version control system that facilitates working on coding and writing projects collaboratively, and allows you to revert your code to a previous version if you realize that you made a mistake. Version control systems such as git are used in most modern data science and statistics positions in industry. Part of my goal as an educator in the statistics program is to ensure that you are prepared to enter the work force, and for that reason the basic use of git is a learning objective for this course. This means that all labs and the computational portion of homework assignments will be distributed to you in git repositories and submitted by committing and pushing the completed assignment to GitHub. I will provide further details and walk through this process, as well as basic interaction with git, in class. Note that we will use the graphical interface to git that is built into RStudio rather than the command line interface to git.

Assignments

Homework Homework is the most effective way to reinforce concepts learned in class. There will be regular homework assignments. Homework assignments will generally include a computational component and a more theoretical component. Late assignments are a headache, and I'd rather not deal with them. I may accept a homework assignment within 48 hours of the due date for a 25% penalty. If you turn in an assignment late, there's a decent chance that it won't be graded until the end of the semester. Extensions may be possible, but need to be requested well before the deadline.

Exams There will be one or two midterm exams and occasional quizzes to be taken during class, as well as a final exam during the exam period. I haven't decided yet whether the exams will be in class, take home, or a mixture of in class and take home.

Writing Your ability to communicate results, which may be technical in nature, to your audience, which is likely to be non-technical, is critical to your success as a data analyst. The assignments in this class will place an emphasis on the clarity of your writing. That said, we are all constantly improving at writing. Your classmates and I are here to help you improve as a writer.

Extra Credit Extra credit is available in several ways: attending an out-of-class lecture (as will be announced) and writing a short review of it; pointing out a substantial mistake in the book, a homework exercise, an exam solution, or something I present in class; drawing my attention to an interesting data set or news article; etc. The extra credit is applied when a student is near the boundary of a letter grade.

Grading When grading your written work, I am looking for solutions that are technically correct and reasoning that is clearly explained. *Numerically correct answers, alone, are not sufficient* on homework, tests or quizzes. Neatness and organization are valued, with brief, clear answers that explain your thinking. If I cannot read or follow your work, I cannot give you full credit for it.

Your grade for this course will be a weighted average of the following components:

Item	Weight
Participation and Labs	10%
Homework	50%
Quizzes and Exams	40%