MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE

Last revised: January 2, 2014

Instructors:

- Paul Bamberg (to be addressed as "Paul," please)

  Office: Science Center 322, (617-49)5-9560

  Email: `bamberg@tiac.net`

  Office Hours:

  Tuesday and Thursdays, 1:30 - 2:20

  Paul graduated from Harvard in 1963 with a degree in physics and received his doctorate in theoretical physics at Oxford in 1967. He taught in the Harvard physics department from 1967 to 1995 and joined the math department in 2001. From 1980 to 2000 he was one of the principals of the speech recognition company Dragon Systems and taught computer science courses in the Extension School. If you count Extension School and Summer School, he has probably taught more courses than anyone else in the history of Harvard. He was the first recipient of the White Prize for excellence in teaching introductory physics and has received the Petra T. Shattuck Excellence in Teaching Award from the Extension School.

  Although Paul's research in computer speech recognition was based on statistical techniques, this course is his first venture into teaching statistics. He is, however, a veteran teacher of probability (MATH E-102 and Mathematics 154) and of computer science (Data Structures, Assembly Language, Theory of Algorithms, GUI programming for Windows and for Linux).

- Jonathan Polit (course assistant)

  Email: `j.l.polit@gmail.com`

  Office Hours:

  Jonathan is a 2010 graduate of Oberlin College with a double major in Psychology and Economics. He has since had the opportunity to apply statistical methods in a wide range of settings, while continuing his mathematical education through the Harvard Extension School. Jon spent a little over a year in Washington, DC, working for a health policy research firm. He then returned to Boston and became involved with the Institute for Quantitative Social Sciences at Harvard, supporting statistical analysis for

a litigation. Jon currently works for Research Computing Services at the Harvard Business School, providing in-house statistical and computational consulting for HBS faculty and students.

Course Website: `http://isites.harvard.edu/k100711`

Goals:

This course has been created to to provide a crucial missing piece for the new Extension School Certificate in Mathematics for Graduate Study. It is designed to give students expertise in using R to analyze data while taking advantage of their background in calculus to prove the key mathematical results that underlie this analysis.

Prerequisites:

- Solid command of single variable calculus and infinite sequences and series. You should be able, for example, to do integration by substitution and integration by parts and to recognize and use geometric series and the Taylor series for the exponential function.
- Some familiarity with multiple integrals. If you have ever taken a course in multivariable calculus, you will be fine. If you took Math E-23a or are enrolled in Math E-23b, that is more than sufficient.
- Familiarity with elementary probability. Although probability will be done "from scratch," it would be nice already to be familiar with concepts like "event," "conditional probability," and "expectation and variance of a random variable."
- Willingness to be held responsible for thirteen proofs. If you have take a course like Math E-23a, you are very well off. Otherwise you will have to work a little harder than your classmates to develop a new skill. Most of the proofs in this course are straightforward algebra.

Course Meetings:

The course meets Wednesday evenings from 7:40 PM to 9:40 PM in Science Center 112. The course assistant will also hold a weekly problem-solving session, whose time will depends on when students in the course are free.

Textbooks:

- Chihara and Hesterberg, Mathematical Statistics with Resampling and R

  This book is available at the Coop. Used copies can be found on amazon.com for as little as $49. The course will follow it quite closely.
- (supplementary)Haigh, Probability Models The book is available electronically through the Harvard library system (use HOLLIS and search for the author and title). If you have not taken a recent probability course, it will be useful as supplementary reading.

Homework: Homework (typically 3 or 4 problems) will be assigned weekly. Many of the problems will require you to use R to analyze the data sets that accompany the textbook.

Homework that is handed in after the day when it is due will not be graded. If it looks fairly complete, you will get a grade of 50% for it.

You are encouraged to discuss the course with other students, and the course staff, *but you must always write your homework solutions out yourself in your own words.* If you find a solution to a problem on the internet or in another book, restate it in your own words and cite your source.

Section problems:

The weekly sections will be used for students, working in small groups, to solve problems, some mathematical, some involving data analysis in R. You should bring your laptop to section every week.

One member of each group should post a solution to the group's problem, either a PDF file or an R script, to the course Web site. Everyone is expected to make four contributions to the site, one for every three classes.

Proofs:

Each class will include a "proof of the week" that you are expected to learn well enough so that you can do it without notes.

Some of the proofs, chosen at random, will appear as questions on the final exam.

On the Web site we will post .pdf files of the proofs written by students in the class.

These are the only proofs for which you are responsible.

LaTeX:

This is the technology that is used to create this syllabus and other course handouts. Once you learn how to use it, you can create professional-looking mathematics on your own computer. There are versions for the PC and the Mac.

We expect that everyone in the class will learn to use LaTeX for their term project and to create .pdf files for some of the problems solved at the end of class.

I learned LaTeX without a book or manual by just taking someone else's files, ripping out all the content, and inserting my own, and so can you. For the PC, you will need to download freeware MiKTeX version 2.8 (see `http://www.miktex.org`), which includes an integrated editor named TeXworks.

From `http://tug.org/texworks/` you can download TeXworks for the Mac OS X.

When in TeXworks, use the Typeset/pdfLaTeX menu item button to create a .pdf file. To learn how to create fractions, sums, vectors, etc., just find an example in the course instructional modules and copy what I did. All the LaTeX source for the modules is on the Web site, so you can find working models for anything that you need to do.

Exams:

There will be a two-hour final exam in class on May 14.

It will include a random selection of three out of the thirteen proofs, a couple of problems based on ones that were solved at the end of class, and a couple of new problems to be solved using R on your own laptop.

Term Project:

You will be expected to apply techniques that you have learned in the course to a data set that is of interest to you, perhaps something from your job or from a hobby. Projects will be presented in class on May 7, and you will be asked to post on the course Web site, in advance, a copy of your data set, an R script that you used to analyze it, and a two-page summary of your analysis and conclusions.

Joint projects with other courses are acceptable, but they must be approved in advance by the instructors in both courses.

Collaboration Policy:

You are encouraged to collaborate with classmates in doing homework, but you must write up solutions in your own words. Students are strictly forbidden to copy one another's R scripts.

You are welcome to discuss your term project with classmates. If you get good ideas from a classmate or from one of the instructors about how to analyze your data set, you should cite your source of advice in your project writeup. Anything that was presented in class, that is in the textbook, or that is in any of the R scripts on the course Web site may be treated as "common knowledge" and need not be cited.

Grades: Your course grade will be determined as follows:

- problem sets, 48 points. Your worst score will be converted to perfect.
- Contributions to the Web site, 12 points.
- term project, 30 points
- Final exam, 60 points.

The grading scheme is as follows:

Points	Minimum Grade
92.0%	A
86.0%	A-
80.0%	B+
74.0%	B
68.0%	B-
62.0%	C+
56.0%	C
50.0%	C-

Schedule:

This schedule is ambitious! It assumes that the authors of the textbook are correct in saying the the book can be covered in a one-semester course.

January 27	Probability and statistics illustrated by data frames in R (Chapter 1)
February 5	Using R to explore data sets (Chapter 2)
February 12	Permutation-test methods for hypothesis testing(Chapter 3)
February 19	Classical methods for analyzing data frames(Chapter 3)
February 26	Sampling distributions and the central-limit theorem(Chapter 4)
March 5	Bootstrapping(Chapter 5)
March 12	Estimation (Chapter 6)
March 19	SPRING BREAK - NO CLASS
March 26	Student t test; confidence intervals (Chapter 7)
April 2	Bootstrap confidence intervals(Chapter 7)
April 9	Classical hypothesis testing(Chapter 8)
April 16	Linear and logistic regression(Chapter 9)
April 23	The Bayesian approach(Chapter 10)
April 30	Smoothed bootstrap; Monte Carlo integration(Chapter 11)
May 7	Project presentations
May 14	Final exam