



The University of California - Los Angeles
Department of Statistics

Introduction to Computational Statistics with R

Statistics 102A

Professor: Mark S. Handcock

Syllabus and Description

Professor:

Mark S. Handcock 8141 Mathematical Sciences Bldg

Motivation and Synopsis

During the twentieth century, the development of statistical computing played a crucial facilitating role for the growth of the statistics discipline and the adoption of statistical methods within the scientific community and beyond. In the twenty-first century digital age, the amounts of data available for statistical analysis has grown tremendously, yielding new opportunities for statistical computing, as well as new challenges. Statistical computing constitutes an important part of a statistics education, and is highly valuable for statisticians in both academia and industry.

This course is designed to provide the upper-division statistics student with the fundamentals of statistical computing, particularly through use of the language R.

The course is thematically split into two parts. The first part will focus on learning the tools and the necessary skills to perform computational statistics. Students will learn intermediate to advanced R programming and usage of some of its functions and packages. The student will learn how to develop functions and packages for the management, pre-processing and analysis of statistical data. The second part of the class will focus on some foundational methods in computational statistics. This includes numerical methods such as root finding, numeric integration, and mathematical optimization. It will continue on to cover the generation of random variables, simulation, and Monte Carlo methods to answer statistical questions.

The computer is the scientific laboratory of the statistician. It plays the same role for the statistical research as the traditional laboratories play for physics and chemistry researchers. As such this course should allow the student to develop a degree of comfort and competence “in the lab.”

The primary purpose of this course is to provide students with a common set of core knowledge about statistical computing for their class work and research. The course will

have an applied focus on tools. The course will involve the practical application of the ideas of statistical computing and their implementation through statistical software, particularly R.

Structure of the Course

There will be a two lectures per week. There is also a discussion section on Thursday.

Textbooks

- [W] Wickham, Hadley.
Advanced R. (2014).
Chapman and Hall/CRC.
Required and available online as a ebook from the library, free for UCLA students.
<http://www.crcnetbase.com/doi/book/10.1201/b17487>
- [C] Chang, Winston.
R Graphics Cookbook. (2012).
O'Reilly.
Required and available online as a ebook from the library, free for UCLA students.
<https://goo.gl/IJM16i>
- [M] Maillardet, Robert, Owen Jones, and Andrew Robinson.
Introduction to Scientific Programming and Simulation Using R. (2014).
Chapman and Hall/CRC.
Required and available online as a ebook from the library, free for UCLA students.
<http://www.crcnetbase.com/doi/book/10.1201/9781420068740>
- [Z] Zieffler, Andrew S., Jeffrey R. Harring, and Jeffrey D. Long
Comparing Groups: Randomization and Bootstrap Methods Using R. (2014).
Wiley.
Required and available online as a ebook from the library, free for UCLA students.
<http://onlinelibrary.wiley.com/book/10.1002/9781118063682>
- [G] Gentle, James E
Matrix Algebra: Theory, Computations, and Applications in Statistics (2007).
Springer
Required and available online as a ebook from the library, free for UCLA students.
<http://link.springer.com/book/10.1007%2F978-0-387-70873-7>

Students must be connected to the UCLA network to obtain their free download. Students who would like to download the textbook off-campus may do so by connecting the the UCLA network via VPN <https://www.bol.ucla.edu/services/vpn/all.html>.

In addition to these books, there are a multitude of books covering pieces of the course content and from varying perspectives. I suggest you use one or more of the following.

You can read the following excellent books online for free from any UCLA account. For computational statistics I suggest

[GH] Geof H. Givens and Jennifer A. Hoeting
Computational Statistics, 2nd Ed. (2013).
Wiley series in computational statistics.
Available online as a ebook from the library, free for UCLA students.
<http://onlinelibrary.wiley.com/book/10.1002/9781118555552>

[JC] Chambers, John M.
Software for Data Analysis: Programming with R. (2008).
Springer: New York.
Required and available online as a ebook from the library, free for UCLA students.
<http://link.springer.com/book/10.1007/978-0-387-75936-4/page/1>

You can read the other books with different perspectives online for free from any UCLA account, starting from:

proquest.safaribooksonline.com/search?q=BOOKTITLE%20r

In particular, I point out:

Joseph Adler
R in a Nutshell

Paul Teetor
R Cookbook

Course Webpage and Discussion Forum

The course has a webpage through the UCLA *Common Collaboration & Learning Environment* (CCLE) system, www.ccle.ucla.edu. The webpage will be continuously updated throughout the course with handouts, homework assignments and solutions. Users sign in to CCLE with their UCLA Logon IDs.

I will be using the Forum feature of the CCLE to provide discussion of issues in class and related questions. For questions that might be of interest to other students, please use the mailing list rather than solely emailing me. Example of questions are about interesting articles you have seen in the media, problems with access to resources, homework or computer questions. Enjoy!

Please regularly consult this class CCLE home page and archive of the mailing list. It will contain lecture notes, homework, solutions and course information.

Computer Usage and Software

The computer is the scientific laboratory of the applied researcher in quantitative fields. As such this course requires the student to develop a degree of comfort and competence “in the lab”.

Our computer interface to R will be the RStudio IDE, which you can download from www.rstudio.com I suggest you look at

John Verzani
Getting Started with RStudio

which you can read online at

proquest.safaribooksonline.com/book/programming/r/9781449314798

Course Requirements and Grades

- 60% Homework (6 assignments, 10% each, none are dropped)
- 15% Midterm Exam (in class, Week 6 Monday, handwritten, 45 minutes)
- 20% Final Exam (scheduled final exam time, handwritten, 2 hours)
- 5% Class Attendance and Discussion forum participation (Attendance is taken on arbitrary days)

Homework

There will be six homeworks that cover both the theory and computation. All assignments involve writing and running R code.

Each homework assignment is worth the same amount. Each homework assignment will be scored on a scale of 20 points. None of the homework scores will be dropped. It is your responsibility to verify that your homework assignment successfully uploaded by the deadline. All homework assignments will be posted as an R Markdown file. Students will download the R Markdown file, modify it with his or her own answers, and submit the rendered HTML file electronically via the CCLE *Homeworks* page. The grader or professor may also request the original `Rmd` file in addition to the `html` file.

Files must not refer to any resources on the local machine or to files that are not publicly available online. No one should make manual edits to a data file on his or her local machine.

Students are free to discuss homework problems and solutions. Discussing the contents of the course with fellow students can be a valuable element of the learning process, and doing so is therefore generally encouraged. However, each student must hand in their own solutions, and the student should, if asked, be able to explain the solutions.

Late Policy for Homework

There is a 10 minute grace period. Submissions up to 10 minutes late will be accepted with no penalty.

Homework assignments submitted 11 minutes late or more will be accepted with penalty. There is a minimum deduction of 10 points for being less than one hour late. An additional 5-point deduction will be taken for each additional hour it is late.

Thus, an assignment that is between 11 and 59 minutes late will receive a 10-point deduction. An assignment that is between 1 hour and 1:59 late will receive a 15-point deduction. An assignment that is between 2 hours and 2:59 late will receive a 20-point deduction, and so on.

I strongly advise uploading the homework to the CCLE well in advance of the deadline in case there are connectivity problems or server issues.

Exams

There will be two written exams, a Midterm Exam (15%) and a Final Exam (20%). While not finally confirmed, the tentative date for the Midterm is Monday October 31st The Final Exam will be during Exam Week, as usual. Both are sit down exams.

Syllabus of the Course

The syllabus of the course will develop on the following weekly schedule. The some later topics may not be reached and we will make choice among them toward the end of the quarter.

Week	Contents discussed
1	Introduction, the R language and eco-system.
2	Data structures and their management.
3	R programming and writing functions
4	Importing data, web scraping, manipulating data with <code>tidyr</code> and <code>dplyr</code>
5	Visualization and graphics (<code>ggplot2</code>)
6	Numerical methods: floating point arithmetic, root finding
7	Numerical methods: basic optimization
8	Random numbers, random variables, and simulation in R
9	Randomization tests, permutation tests and bootstrapping
10	Additional topics: Monte Carlo-integration, kernel density estimation

Academic Integrity

As a student and member of the University community, you are here to get an education and are, therefore, expected to demonstrate integrity in your academic endeavors. All students must uphold University of California Standards of Student Conduct as administered by the Office of the Dean of Students. Students are subject to disciplinary action for several types of misconduct, including but not limited to: cheating, multiple submissions, plagiarism, prohibited collaboration, facilitating academic dishonesty, or knowingly furnishing false information. You may have assignments or projects in which you work with a partner or with a group. For example, you are welcome, and even encouraged, to work with others to solve homework problems. Even though you are working together, the assignment you submit for a grade must be IN YOUR OWN WORDS, unless you receive specific instructions to the contrary. For more information about academic integrity, please go to www.deanofstudents.ucla.edu.

I welcome comments or suggestions about the course at any time, either in person, by letter, or by email. Please feel free to use these ways make comments to me about any aspect of the course.

Support

Title IX prohibits gender discrimination, including sexual harassment, domestic and dating violence, sexual assault, and stalking. Students who have experienced sexual harassment or sexual violence can receive confidential support and advocacy at the CARE Advocacy

Office for Sexual and Gender-Based Violence, 1st Floor Wooden Center West, CAREadvocate@caps.ucla.edu, (310) 206-2465. You can also report sexual violence or sexual harassment directly to the University's Title IX Coordinator, Kathleen Salvaty, 2241 Murphy Hall, titleix@conet.ucla.edu, (310) 206-3417.

I welcome comments or suggestions about the course at any time, either in person, by letter, or by email. Please feel free to use these ways make comments to me about any aspect of the course.