**Statistical Computing and Introduction to Data Science (GR5206)**
**Fall 2017**

## Description

Solid programming skills and good computational understanding are necessities for current statisticians making statistical computing an essential element of modern statistics curricula. Statisticians are routinely expected to gather data from disparate sources and implement the most current methodologies, both of which require computational fluency. This course is an introduction to the basics of statistical programming, targeted at entering statistics MA students with minimal prior programming knowledge. Examples from data science will be used throughout the course for demonstration. Students will be introduced to basic machine learning topics such as classification, regression, and clustering methods, resampling techniques including the bootstrap, cross-validation, and permutation tests, as well as the basics of optimization. At the end of the semester students will have:

- The ability to read and write code for statistical data analysis,
- An understanding of programming topics such as functions, object, data structures, debugging, etc.,
- The skills to display analysis results in reproducible ways.

The class will be taught in the `R` language using the `RStudio` interface.

## Administrative

### Lecture

- Time: Friday 2:40pm-5:25pm
- Location: 417 International Affairs Building
- Lab: Meets roughly weekly beginning the week of September 11-15

### Instructor

Cynthia Rush
Office Hours: Tuesday, 5:00pm-6:00pm
Main Office: Department of Statistics, Room 1009, 10th Floor School of Social Work (SSW)
Email: cgr2130@columbia.edu

### Teaching Assistants

- Tim Jones, Section 002
  Email: tdj2113@columbia.edu
  Section Time: MW 6:10pm-7:25pm
  Section Location: 644 Seeley W. Mudd
  Office Hours: Wednesday, 5:00pm-6:00pm
  Location: Statistics Lounge, SSW
- Chaoyu Yuan, Section 003
  Email: cy2438@columbia.edu
  Section Time: TR 11:40am-12:55pm
  Section Location: 214 Pupin Laboratories
  Office Hours: Wednesday, 1:00pm-2:00pm
  Location: Room 1023, SSW

- Shun Xu, Section 004
  Email: sx2220@columbia.edu
  Section Time: MW 8:40am-9:55am
  Section Location: 413 Kent Hall
  Office Hours: Monday, 3:00pm-4:00pm
  Location: Statistics Lounge, SSW

- Fan Gao, Section 005
  Email: gao.fan@columbia.edu
  Section Time: TR 8:40am-9:55am
  Section Location: 413 Kent Hall
  Office Hours: Wednesday, 4:00pm-5:00pm
  Location: Statistics Lounge, SSW

## Prerequisites

STAT GR5204 and STAT GR5205 or the equivalent. Students will also be expected to have basic knowledge of linear algebra, elementary probability, and multivariate calculus.

## Grading and Academic Integrity

Columbia's intellectual community relies on academic integrity and responsibility as the cornerstone of its work. Graduate students are expected to exhibit the highest level of personal and academic honesty as they engage in scholarly discourse and research. In practical terms, you must be responsible for the full and accurate attribution of the ideas of others in all of your research papers and projects; you must be honest when taking your examinations; you must always submit your own work and not that of another student, scholar, or internet source. Graduate students are responsible for knowing and correctly utilizing referencing and bibliographical guidelines. When in doubt, consult your professor. Citation and plagiarism-prevention resources can be found at the GSAS page on Academic Integrity and Responsible Conduct of Research (`http://gsas.columbia.edu/academic-integrity`).

Failure to observe these rules of conduct will have serious academic consequences, up to and including dismissal from the university. If a faculty member suspects a breach of academic honesty, appropriate investigative and disciplinary action will be taken following Dean's Discipline procedures (`http://gsas.columbia.edu/content/disciplinary-procedures`).
Your grade will be determined by three different components:

- **Homework (20%).**
- **Lab (10%).**
- **Midterm Exam (35%).**
- **Final Exam (35%).**
- **Course Participation.** After the course grades are calculated and curved, I will increase the letter grade of the most active participants in the course. Course participation primarily includes participation in lectures.

Failure to complete any of the first three components may result in a D or F.

**Homework:** `R` and `RStudio` will be used throughout the course and for homework assignments. Homeworks will be due weekly. All homework must be turned in online through the Canvas page in PDF format, have a .pdf extension (not zip or other archive), and be less than 4MB. To receive full credit, students must thoroughly explain how they arrived at their solutions and include the following information on their homeworks: name, UNI, homework number (e.g., HW03), and class (STAT GR5206). The assignments will be posted on the Canvas page every Monday and due 8pm

the next Monday unless stated otherwise. **Late homework will receive a grade of zero (see the late work policy below). To compensate, the lowest homework score will be dropped.** You are encouraged to work with other students on the homework problems, however, verbatim copying of homework is absolutely forbidden. Homework write-ups must be done individually and must be entirely the author's own work. Homeworks not adhering to these requirements will receive no credit.

**Labs:** Weekly labs will begin the week of September 11-15. During each lab session, students are encouraged to work in groups on a small in-class project using `R`. The lab sessions will help solidify the concepts covered during lecture. The labs should ideally be completed during the scheduled lab hours. The course TA will be available to assist students during the lab. If students do not finish the worksheet in the scheduled session, they must submit the lab report by 8:00pm that day. Attendance is required in these lab sessions. The labs contribute 10% of students final grade which accounts for both attendance and the completed lab.

**Late Work and Regrading Policy:** No late work or requests for regrades are accepted. To accommodate unexpected circumstances, we have implemented two important features:

- Your lowest homework grade will be automatically dropped at the end of the term.
- You may submit and resubmit your homework as many times as you like up until the deadline. This means that you should submit any partial solutions as you complete them, to make sure you receive as much credit as possible for the work you have done. After the deadline, the system will not allow you to submit your homework. If you do not submit anything by the deadline, you will get a 0. **There will be no exceptions to this rule. Submit your homework early.**

**Exams:** There will be an in-class midterm exam and a final exam. Make-up exams will not be given routinely. If you have a legitimate conflict with an exam date, it is incumbent upon you to make arrangements with the instructor to take the test early. An exam missed due to a documented illness or other unforeseeable (and documented) extraordinary circumstances must be made up before the test papers are returned to the class.

## Reading Material

The following resources will be useful at different points in the course and we'll point to them throughout, though no explicit readings will be assigned. In general, a Google search can turn up many good resources about using `R`, and otherwise I've listed a few here.

- Garrett Grolemund and Hadley Wickham, R for Data Science.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning.
- Roger Peng, R Programming for Data Science.
- Cosma Shalizi, Advanced Data Analysis from an Elementary Point of View.
- W.N. Venables, D.M. Smith, and the R Core Team, An Introduction to R.

Finally, this course was developed for Columbia students with much guidance from the following courses. Their web pages are also a good resource for students.

- Jeff Goldsmith (2017), "Data Science 1".
- Cosma Shalizi and Andrew Thomas (2014), "Statistical Computing 36-350: Beginning to Advanced Techniques in R".
- Chris Paciorek (2015), "Statistics 243: Introduction to Statistical Computing".

**Tentative Outline**

| Section | Content |
|---|---|
| 1 | Introduction to `R` and `R`Studio: scripts and Markdown |
| 2 | `R` Basics: data types and best practices. |
| 3 | `R` Basics: EDA and base `R` graphics. |
| 4 | `R` Basics: web scraping and text data. |
| 5 | `R` Basics: writing functions and reproducibility. |
| 6 | `R` Basics: `R` commands for parametric distributions and random number generation. |
| 7 | Introduction to the tidyverse |
| 8 | Midterm |
| 9 | Visualization and advanced `R` Graphics |
| 10 | Split/Apply/Combine and `plyr` |
| 11 | Debugging techniques |
| ... | (the following topics, as time allows) |
| 12 | Databases and parallel processing |
| 13 | Packages and GitHub |
| | Final Exam |