Statistics GU4206/GR5206
Statistical Computing and Introduction to Data Science
Fall 2019

Instructor
Gabriel Young
Email: gjy2107@columbia.edu
Office: Room 614, 6th floor of Watson Hall, 612 West 115th Street

Teaching Assistants

- **In-class:** Owen Ward
  Email: owen.ward@columbia.edu
- **Online:** Fan Gao
  Email: gao.fan@columbia.edu

Office Hours

- Gabriel Young:
  - On campus students: Mondays and Wednesdays 11:30am-12:30pm, room 614 Watson Hall and Watson Lounge.
  - Online students: Appointment basis
- Owen Ward:
  - On campus students: Mondays and Wednesdays 3:00pm-4:00pm, SSW 10th floor lounge.
  - **The in-class TA will not be available for online students**
- Fan Gao:
  - Online students: Mondays 8:30pm - 10:30pm, Zoom
  - **The online TA will not be available for in-class students**

Learning Outcomes
Statistical computing is an essential element of modern statistics curricula as solid programming skills and good computational understanding are necessities for current statisticians. 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 master

students and senior undergraduate 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, 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.,
- An introduction to statistical learning methods applied to real-word data.

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

Class Time and Location
STAT GU4206/GR5206 is a 3 credit hour course. The class meets Fridays from 8:40am - 11:25am, 903 School of Social Work.

Prerequisites for GR4206
STAT GR4204 and GR4205 or the equivalent. Students will also be expected to have basic knowledge of linear algebra, elementary probability, and multivariate calculus.

Prerequisites for GR5206
STAT GR5204 and GR5205 or the equivalent. Students will also be expected to have basic knowledge of linear algebra, elementary probability, and multivariate calculus.

Text/Supplies
Note that all text books are optional. I highly recommend the the first three.

- *R for Data Science*; Garrett Grolemund and Hadley Wickham.
- *An Introduction to Statistical Learning*; Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
- *Computational Statistics*; Geof Givens and Jennifer Hoeting.
- *Advanced Data Analysis from an Elementary Point of View*; Cosma Shalizi.
- *The Art of R Programming: A Tour of Statistical Software Design;* Norman Matloff.

- Maybe more..

Software
R and RStudio will be used throughout the course and the assignments. R is open-source statistical software that can be downloaded at https://www.r-project.org and RStudio at https://www.rstudio.com. We expect that students will have the software downloaded before class begins.

Homework
Homework assignments will be posted on Canvas at least one week in advance of the due date. There will be 6-8 assignments given during the semester. All homework must be created using **R markdown** and turned in online through the Canvas page in PDF or HTML format. The PDF files must 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). Late homework will receive a grade of zero (see the late work policy below). 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
Biweekly labs will begin the second week of class. 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 are designed to be completed during the scheduled lab hours. If students do not finish the worksheet in the scheduled session, they must submit the lab report as a pdf or html by the deadline. Attendance is required during these lab sessions. The lowest lab score will be dropped.

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 lab 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 so-

lutions 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.

Class Structure
The class follows a traditional lecture environment. New topics are presented in each class. In class examples and other relevant course materials will be posted on Canvas. The slides are covered during the scheduled contact hours.

Attendance
Students should ideally attend each class meeting. Even though attendance is not required, I will frequently give examples or hints during lecture that may show up on assessments.

Grading

Homework	20%
Lab	10%
Midterm	35%
Final	35%

Exams
You will have 150 minutes to complete the midterm exam, and 170 or 180 minutes to complete the final. Allowable materials will be discussed one week before the exam.

Midterm
There will be one midterm exam. The midterm exam is scheduled for

- 10/18/2019

Final
The final exam will be weighted 30% of the final course grade. The final exam is tentatively scheduled for

- TBA

You must take the final at the scheduled time.

Exam absences
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.

Academic Honesty
The university expressly prohibits academic dishonesty such as cheating, plagiarism, etc. It provides for a number of rather unpleasant consequences for students who are caught in violation of its academic honesty policies. Any suspected cheating on examinations will be referred to the Dean's Discipline process, possibly resulting in course failure or College dismissal.

Grading Scale

93 or more	A
90 to 92	A-
87 to 89	B+
83 to 86	B
80 to 82	B-
77 to 79	C+
70 to 76	C
60 to 69	D
Below 59	F

The letter grade $A+$ will be given to the top two students in the class (assuming they earned a final grade of at least 93%).

Tentative course outline

Date	Topic	Lab week
09/06/2019	Introduction to R and RStudio. Working with data in R.	
09/13/2019	Working with data in R continued including: data frames, iterative coding.	YES
09/20/2019	R base graphics. Linear algebra review. Multiple linear regression, Bootstrap procedure	
09/27/2019	Character strings. Regular expressions. Web scrapping	YES
10/04/2019	Writing functions. Basic classification methods	
10/11/2019	Split/Apply/Combine.	YES
10/18/2019	Midterm	
10/25/2019	Tidyverse and ggplot	YES
11/01/2019	Random number generation. Simulation. Monte Carlo integration.	
11/08/2019	Simulation continued.	YES
11/15/2019	Distributions as models. MCMC	
11/22/2019	Optimization. Logistic regression.	YES
11/29/2019	Fall break	
12/06/2019	Other Topics	
12/13/2019	Final exam week (Final date TBA)	