Department of Statistics
The Wharton School
University of Pennsylvania
400 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104.6340
215.898.8222 phone
215.898.1280 fax

Wharton
University of Pennsylvania

Statistics 470/503/770: Data Analytics and Statistical Computing Syllabus Fall 2018

Time and Location

- Section 1: Mon/Wed, 12noon-1:20pm, 351 SH-DH
- Section 2: Mon/Wed, 3:00-4:20pm, 351 SH-DH

General Content:

- This course offers a gentle introduction to programming for non-programmers in the widely used R language.
- In addition, the course presents basics of text analysis, data visualization, simulation, and statistical inference based on simulation (bootstrap).

Enrolling:

- If you need special permissions for this course, **do not contact the instructor** but the statistics course administrators at `stat-courses@wharton.upenn.edu`.
  Please, make sure you actually need a permission.
- This course is cross-listed as Statistics 503 to count as 500-level for graduate students outside of the Statistics department. It is also cross-listed as Stat 770 for MBAs.
- The next offering of the course is likely to be in the fall of 2019.

Organization:

- Instructor: A. Buja
  Office hours: Mon 4:30-6:30pm, 471 or 440 JMHH
- TA: Linjun Zhang (email: linjunz – at wharton...)
  Office hours: Thu 4:30-6:30pm, Room JMHH 260
- Online Canvas will be explained in the second class. It will be used for announcements, class notes, homeworks, solutions, and discussions.

Prerequisites:

- Stat 111/112 or Stat 101/102 or Stat 431 or Stat 613.
- No programming experience is required. In fact, you **should** have no such experience. Experienced programmers: Please, **do not take this course!**.
  There are more efficient ways for you to pick up R (e.g., youtube and Coursera offerings).

Please, free up your slot for someone who needs an introduction to programming. You represent unfair grade competition if you work along-side students who have yet to get used to the rigors of strict syntax, the hardships of debugging code, and mental models of computation.

Class Materials:

- Textbook: None
- Required readings: to be provided online.
  - The chapter files on Canvas under 'Files' > 'Chapters':
    These files contain explanatory text, R code examples, and unsolved coding problems. The latter will be solved in class. Solutions will **not** be posted! If you missed class or were too slow to follow, obtain solutions from a friend.
  - The file 'Recaps-and-Comments.R' on Canvas under 'Files':
    This file will be updated for each class with with organizational announcements as well as recaps and extensions of the previous class. It will be used at the beginning of each class.
  - This syllabus: Before you ask organizational questions, make sure they aren't already answered here.
- Software: "R" as packaged by "RStudio" at
  `https://www.rstudio.com/products/rstudio/download/`
  Look under "Installers for Supported Platforms."
  You can also download a "cheat sheet" with concise instructions for RStudio.
- MAC Users: If RStudio does not work for you, you may also have to download the raw version of R from `https://cran.r-project.org`
- The following materials are NOT required, but may serve those students who like to read supplemental materials:
  - Go to https://cran.r-project.org/, click "Documentation" > "Contributed".
  - Search the internet for introductions to R; they exist in large numbers, some free, some for purchase.

Organizational Details and Procedures:

- **Grades:**
  - They will be based on (1) homeworks, (2) in-class quizzes, and (3) brownie points based on three forms of participation.
  - "Brownie points" can be acquired from participation (a) in class, (b) in the Canvas discussion board, and (c) by attending office hours. Brownie points **may** give you a "bump" from, say, B+ to A−, if your overall score is near the upper boundary of the letter grade bracket.
  - There will be no midterm and no final exams, but the last homework will be at the end of the semester and overlap with the reading and final exam periods.
  - The overall score from homeworks and quizzes will be "curved" because there is no way for an instructor to know how difficult his homework problems and quiz questions are.
- Inquiries about grades and scoring of homeworks and quizzes should be directed at the TA, not the instructor. Use Canvas email or TA office hours.

- **Quizzes** are in-class, multiple choice, 12 questions each.
  Permitted: 1 cheatsheet, letter size, both sides, printed or hand-written, created by you, not verbatim copied from other students or other sources.
- **Homework conduct and collaboration:**
  1. To promote learning, you are encouraged to work with each other, including discussing homeworks.
  2. **You must, however, submit work that you created on your own — AFTER discussing homework with others. There is a strict prohibition of copying others' solutions.**
  3. Graders will report cases of copying, and there will be considerable point deductions depending on the seriousness of the case.
- There is no schedule for homeworks and quizzes yet because the material of the class will be reworked to some extent.
- Important for homework and quizzes is that you develop skills of reasoning about code using slow and systematic thinking. You will not be able to "feel" yourself to correct solutions.
- **Stress relief:** If you are unable to complete a homework by the deadline or participate in a quiz at the announced time, you will have a budget of late submission days for homeworks and a budget of make-up quizzes you can draw on. The rules are as follows:
  - The budget for late submission of homeworks is a total of 7 days for the semester. For example, if you submitted three homeworks that were late by 2 days, 3 days and 2 days, you have reached the budget limit and subsequent homeworks need to be on time. Else there will be a deduction of 4 points (usually out of 20) for each additional late day. It is important that you keep track of your late homework budget and not let any homework slip indefinitely! Canvas time-stamps your uploaded solutions.
  - There will be an announced date (usually TA office hours) for the make-up quiz, 2-4 days after the in-class date. The budget for taking make-up quizzes is 2. If by the end of the semester there are more than 2 quizzes taken during make-up time, there will be a deduction of 3 points (usually out of 12) for each quiz over budget. Exceptions can be granted for good cause but need to be approved by the instructor beforehand.
  - If you find yourself stressed out, please, do make one more effort and talk to the instructor and/or the TA to find some kind of accommodation that will bridge you over. Importantly, talk before it is too late! Make sure, though, that your stresses are not due to inappropriate behaviors or priorities.
- **Class room rules of conduct**:
  1. Respect and helpfulness to fellow students.
  2. No disruption of learning:
     - If you must leave early, let the instructor know and sit near an exit.
     - Same if you expect an urgent phone call.
     - Minimize flickering on your computer screen to avoid distraction of your fellow students' peripheral vision.
     - No food in the class room. Especially no smelly foods! Beverages permitted: water, coffee, tea; but, please, do not snap open soft drink cans during class.

3. No electronic media in the class room (texting, email, facebook, instagram, browsing, games, working on extraneous tasks, ...) except when related to the class.
4. Professional expression in all forms of communication (oral, email, written), free of slang.
   In particular, there is a ban on the use of "**like**" as a filler word.
   Use our conversations as practice for your job interviews where you want avoid the impression of immaturity.
5. Honesty and openess in all proceedings, including homeworks and quizzes.

- **Computers in class**: You are encouraged to follow the class actively by scrolling down the current chapter file in RStudio and writing, executing and debugging code in the file as we go along. However, you are under honor code not to use the computer for purposes other than class-related matters.
- **Experimental nature of the course**: This course is being adjusted as we go along. We may not be able to issue as many homeworks and quizzes as would be optimal.
- **Facing programming the first time**: We expect a very heterogeneous group of students in this class. This course is meant to get you started if you have never programmed. Students for whom the pace seems slow, but who insist on taking the course anyway, are asked to be patient and helpful to students with lesser background. Students who struggle initially with the exacting demands of programming are asked to look for help from other students, the TA, the instructor, and even outside sources (google searches, youtube videos, ...). It is recommended not to get bogged down in debugging a programming error; instead, look for help quickly so you can move on to the next step.

Emphasis of this Course:

This course emphasizes deep thinking about programming concepts underlying the R language. We will ask: How does the language work? What are its objects? What is the effect of executing R functions? To this end you will need to understand some seemingly arcane aspects of any computer language: (1) Syntax — what expressions are legal in the language? (2) Semantics — what do the expressions mean and what do they compute? (3) How do we make use of these principles to achieve computational solutions to data analytic problems? So, please, do not expect whiz-bang and fireworks. This class builds up slowly but steadily. In the second half of the semester you will start seeing some surprising powers of computation.

Another emphasis, later in the course, will be on statistical principles that will allow us to see statistics in a new light. As we explore simulation-based analysis methods, in particular the bootstrap method, we will come to a fuller understanding of statistical inference. We will also gain access to very general principles that allow us to produce statistical inference for most data analysis situations you may encounter in the future.

Course Content:

Only a selection of the following will be covered. It is more important to understand few concepts, but deeply, than lots of functionality, but superficially.

1. R programming, interwoven with other materials:
   (a) Syntax
   (b) Atomic data types: numeric, character, logical, missing
   (c) Data structures: vectors, matrices, arrays, dataframes, lists

(d) Coercion from one data type to another
(e) Data queries: enumeration, exclusion, Boolean selection, associative arrays
(f) Control constructs: loops, conditionals
(g) Vectorization for efficient computing
(h) Programming principles:
    i. Expressions and evaluation in the functional paradigm
    ii. Names and namespaces
    iii. Functions as objects, their arguments and execution
    iv. Class-based object-oriented programming and generic functions
    v. Pipes

2. Input/output for
    (a) Sequential data
    (b) Data tables
    (c) Web pages
3. Model language and model formulae as used in
    (a) Statistical models
    (b) Plotting
4. Text analysis and language tools:
    (a) Text pasting, splitting, substituting, extracting, formatting
    (b) Regular expressions: text patterns, regexp syntax
    (c) Document-by-term matrices
    (d) Web page scrubbing
5. Data janitorial work:
    (a) Data cleaning
    (b) Feature extraction
    (c) Joining multiple data tables
    (d) Missing values; simple imputations
6. Data visualization, fundamental techniques:
    (a) Univariate plots for all variable types
    (b) Bivariate plots for all combinations of variable types
    (c) Coplots for higher-dimensional data
7. Stochastic simulation:
    (a) Probability distributions
    (b) Random numbers
    (c) Simulating draws from simple probability models
    (d) Simulating draws from regression models
    (e) Random sampling of finite populations
    (f) Accuracy of simulation estimates

8. Nonparametric statistical inference based on simulation:
    (a) Fundamentals of statistical inference:
        i. The real meaning of statistical uncertainty
        ii. Standard errors
        iii. Statistical tests
        iv. Confidence intervals
    (b) Standard errors for almost anything from nonparametric bootstrap
    (c) Visual inference with the line-up method