90-711 Statistical Reasoning with R

Fall 2021

Professor: Dr. Amelia M. Haviland (haviland@cmu.edu)

Head TA: Alec McClean (amcclean@andrew.cmu.edu)
Lab Director: Max Rubinstein (mrubinst@andrew.cmu.edu)
TA Team:
Ernest Afflu (eafflu@andrew.cmu.edu),
Lydia Barit (lbarit@andrew.cmu.edu),
Yuxin Du (yuxind2@andrew.cmu.edu),
Alec Harkins (aharkins@andrew.cmu.edu),
Fahmi Islami (fislami@andrew.cmu.edu),
Bella Lu (yijinl@andrew.cmu.edu),
Ekene Ohanwusi (eohanwus@andrew.cmu.edu),
Wanyi Weng (wanyiw@andrew.cmu.edu)

Statistical reasoning is essential to learning from data and understanding the strengths and limitations of data analyses. This course is grounded in questions of importance in public policy and management and focuses on how data and statistical reasoning can inform those questions. We will use a hands-on approach to develop skills and critical thinking in the fundamentals of causal inference, univariate and bivariate descriptive statistics, quantifying uncertainty, statistical inference, and linear regression. The hands-on approach involves learning the basics of how to use the R statistical language and weekly labs in which students will use R to carry out data analysis on real-world policy issues in a supervised setting. While useful, no R or computer programming experience is required for the course, and this course does not replace the stand-alone R course.

This is a rigorous graduate school introductory statistics and data analysis course, and students who take this course will be enabled to use (and further develop) statistical reasoning and methods in your other courses, your summer internship, your Systems Synthesis Project, and in your career. In today's information world, data are available everywhere and the role of statistics is rapidly increasing in public policy, information security, health care, the arts, the entertainment industry, business, academia and many other parts of society. We echo the message of *The New York Times* which published an article entitled "For Today's Graduate, Just One Word: Statistics."

Format of the Class:

This class meets In Person. It has two sections:

- Section A meets 1:25-2:45pm ET Monday and Wednesday with Lab on Fridays at the same time.
- Section B meets 3:05-4:25pm ET Monday and Wednesday with Lab on Fridays at the same time.
- Lab meetings and some office hours may be online rather than in person – more information to come.

If the class needs to move online, you will receive an email from me (the instructor), and an announcement will be published on our course website on Canvas.

Required Text:

Imai, Kosuke (2018). Quantitative Social Science: An Introduction. Princeton University Press. Paperback ISBN 978-0-691-175461 or eBook ISBN: 978-1-400-885251

It is available in hardcover, paperback, and e-Book. Chapter 1 will be available on the Canvas site for the class. The chapters of the book that we use in this class are available for 2 hour check-out at the library – it is an electronic version from scans of the hardcopy.

Required Software:

In this course we will use the open-source statistical software R (http://www.r-project.org). R can be more powerful than other statistical software such as STATA or SAS, but it can also be more difficult to learn. A variety of resources will be made available for 90-711 students in order to learn R as efficiently as possible. To help make using R easier, we'll be using RStudio (http://www.rstudio.com/)—a user interface that simplifies many common operations and creates professional quality reports of data analyses. Please see https://swirlstats.com/students.html for guidance on downloading and installing these programs to your computer. You will need to complete this prior to the first lab (Friday, September 3).

Learning Objectives: Use statistical reasoning to learn from data

- Use R and R Studio to explore, summarize, and visualize data.
- Apply the concept of potential outcomes to evaluate estimates of causal effects.
- Summarize and interpret univariate and bivariate distributions using histograms, box plots, bar plots, and scatter plots.
- Perform linear regression with single or multiple predictors and assess model fit.
- Interpret the results of linear regression models.
- Use probability to quantify uncertainty in estimators of parameters of interest.
- Make accurate statistical inferences using confidence intervals and standard errors.
- Appropriately interpret results of data analyses and statistical inferences.
- Create reports of data analyses and interpretations of results using R Markdown.

Grade Components:

15% Weekly Labs – All students are required to synchronously attend the weekly lab sessions. In these labs you will work with one or two other students to complete - in a supervised setting - a data analysis that uses the concepts covered in class that week and the R skills covered in the textbook and supplementary materials up to that point. In general, the labs will be more difficult than but otherwise similar to the HW assigned the same week. Your lab with the lowest grade each mini will be dropped (i.e. not included in the calculation of your course grade). You may

complete up to two labs per mini asynchronously if you are not able to attend lab due to illness or other pressing conflicts. Details: grade dropping will occur as stated unless the lowest lab grade occurs the same week as the lowest homework grade, in which case we will drop the second lowest lab grade.

30% Weekly Homework - For the HW and labs, students are required to create a report of their work, code, results, and discussion, using Rmarkdown. This Rmarkdown file in turn will be used to create the report that you will submit for a grade – a final PDF document. The PDF will be uploaded for grading through an application called GradeScope. Instructions on how to do so will be provided. Students may discuss the HW assignments in general, but all code, results, and interpretation of results are to be done individually. The homework assignment with the lowest grade each mini will be dropped (i.e. not included in the calculation of your course grade). *No late homework assignments will be accepted.*

5% Success Kit Mastery Quizzes – This class focuses on higher level concepts. To make sure all students are ready to learn those higher level concepts, asynchronous learning materials will be provided with associated **required** online quizzes to assess your mastery of the 'level one' required concepts, or to provide more practice. The quizzes will be on the course's Canvas site. Students are allowed up to three attempts at these short quizzes prior to a deadline. You will receive credit for completing the quizzes and not completing them will count against you. Your score does not matter, you receive full credit for taking it, regardless of your score. It is up to you to take appropriate steps to learn 'level one' concepts using the provided materials if your quiz score indicates to you that you need to.

12.5% each (*25%* total) In Class Exams – There will be two in class exams held during class time or labs during the term.

25% Final Exam – There will be a 3-hour cumulative final exam held during the exam period at the end of the semester.

Student **active learning** responsibility:

We are partners in this learning experience. I EXPECT YOU TO:

- Conduct your learning with academic integrity, see below for specifics and definitions
- Attend and constructively participate in class and labs
- Read assigned textbook sections, take Mastery Quizzes, and address gaps in Level One skills **before** class
- Do individual homework assignments
- Prepare for exams
- Be aware of and proactive about your own learning style and time management
- ***Pursue*** your own understanding: What is your understanding? Does it fit with what else you know? What is solid, missing, or vague? What you can do to make what is missing or vague more solid: attend office hours, participate in Canvas Discussions, create a study group, review reading, review homework solutions, or make an appointment with the instructor or a TA...

Use of Technology During Class:

This semester involves regular use of technology during lab and occasionally during class time. Research has shown that divided attention is detrimental to learning, so I ask that you close any windows not directly related to what we are doing while you are in lab and refrain from using any technology during class unless specifically instructed to do so. Please turn off your phone notifications and limit other likely sources of technology disruption, so you can fully engage with the material, each other, and the instructor. This will create a better learning environment for everyone. If your technology use is distracting other students I will intervene (ask you to change what you are doing or leave the class/lab) in order to maintain a positive learning environment.

Statement Regarding Diversity and Inclusion

We must treat every individual with respect. We at Heinz College are diverse in many ways, and this diversity is fundamental to building and maintaining an equitable and inclusive campus community. Diversity can refer to multiple ways that we identify ourselves, including but not limited to race, color, national origin, language, sex, disability, age, sexual orientation, gender identity, religion, creed, ancestry, belief, veteran status, or genetic information. Each of these diverse identities, along with many others not mentioned here, shape the perspectives our students, faculty, and staff bring to our campus. We, at CMU, and I as your professor, will work to promote diversity, equity and inclusion not only because diversity fuels excellence and innovation, but because we want to pursue justice. We acknowledge our imperfections while we also fully commit to the work, inside and outside of our classrooms, of building and sustaining a campus community that increasingly embraces these core values. Each of us is responsible for creating a safer, more inclusive environment. It is my intent that students from all diverse backgrounds and perspectives be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. To enable us to address each other respectfully and accurately, please record the name you would like us to use for you and your gender pronouns in NameCoach on Canvas. You can record your name in Canvas --> Account --> Settings --> Name Recording and Pronouns. Please record your name and pronouns promptly.

Accommodations for Students with Disabilities:

If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Statement of Support for Students' Health & Well-being

Your graduate school experience might prove to be exciting, stimulating, and enjoyable, but it is likely to entail stress as well, particularly under ongoing pandemic conditions. The University Provost provides the following thoughts for students. I consider them important.

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. There are many helpful resources available through CMU and an important part of the graduate school experience is learning how to ask for help. Asking for support sooner rather than later is almost always helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

Academic Integrity:

Students are expected to honor the letter and the spirit of the Carnegie Mellon University Policy on Cheating and Plagiarism. All activities cited in that policy will be treated as cheating in this course. Students are expected to familiarize themselves with this policy. In particular, making any use of prior year HW solutions is an academic integrity violation. Students are also encouraged to review the Carnegie Mellon University Academic Disciplinary Actions Overview for Graduate Students, which details penalties and sanctions, as well as students' rights. I will take a zero-tolerance policy on cheating and plagiarism and will consult with College leadership on appropriate action for all instances of cheating and plagiarism. As the aforementioned policies indicate, penalties can include course failure, suspension, and dismissal from the program.
Carnegie Mellon University Policy on Cheating and Plagiarism
http://www.cmu.edu/policies/student-and-student-life/academic-integrity.html
Carnegie Mellon University Academic Disciplinary Actions Overview for Graduate Students
http://www.cmu.edu/academic-integrity/documents/academic-disciplinary-actions-overview-for-graduate-students.2013.pdf

What constitutes plagiarism in a coding class?[1]

The course collaboration policy allows you to discuss the HW problems with other students, but requires that you complete the HW on your own. You may not refer to another student's code, or a "common set of code" while writing your own code. You may, of course, copy/modify lines of code that you saw in lecture, lab, or the course discussion board.
The following discussion of code copying is taken from the Computer Science and Engineering Department at the University of Washington. You may find this discussion helpful in understanding the bounds of the collaboration policy.

[1] This section is taken from Alexandra Chouldechova's class "Programming R for Analytics."

‘[It is] important to make sure that the assistance you receive consists of general advice that does not cross the boundary into using code or answers written by someone else. It is fine to discuss ideas and strategies, but you should be careful to write your programs on your own.
You must not share actual program code with other students. In particular, you should not ask anyone to give you a copy of their code or, conversely, give your code to another student who asks you for it; nor should you post your solutions on the web, in public repositories, or any other publicly accessible place. Similarly, you should not discuss your algorithmic strategies to such an extent that you and your collaborators end up turning in [essentially] the same code. Discuss ideas together, but do the coding on your own.

Expected Semester Schedule – subject to change:

Mini I

- Introduction & Causality – Week 1
    - Topic: Overview & Potential Outcomes • Chapters 1, 2.1-2.3
    - HW #1 posted on Wednesday 9/1; Due by noon Wednesday, 9/8
- Causality –Week 2
    - Topic: Randomized Studies • Textbook: Chapter 2.4-2.5
    - HW #2 posted on Wednesday 9/8; Due by noon Wednesday, 9/15
- Causality –Week 3
    - Topic: Observational Studies & Describing Variability • Chapter 2.6
    - HW #3 posted on Wednesday 9/15; Due by noon Wednesday, 9/22
- Measurement – Week 4
    - Topic: Plotting to Learn about Univariate Distributions • Chapter 3.3, 3.4
    - Review for Exam One
- Measurement – In Class Small Exam 1 Week 5
    - Topic: Describing Bivariate Relationships • Chapter 3.3, 3.6, 4.2
    - HW #4 posted by Wednesday 9/29; Due by noon Wednesday, 10/6
- Prediction -Week 6
    - Topic: Linear Regression with 1 predictor • Chapter 4.2, 4.3
    - HW #5 posted on Wednesday 10/6; Due by noon Wednesday, 10/13
- Prediction – Week 7
    - Topic: Multiple Linear Regression • Chapter 4.3, 3.4, 6.3.1
    - HW #6 posted on Wednesday 10/13; Due by noon Wednesday, 10/20

Mini II

- Probability – Week 8
  - Descriptive versus Inferential Statistics & Probability • Chapter 6.1, 6.2
  - Review for Small Exam 2
- Linear Regression – In Class Small Exam 2 – Week 9
  - Statistical Inference, Random Probability Sampling • Chapter 6.1-6.3
  - HW #7 posted on Wednesday 10/27; Due by noon Wednesday, 11/3
- Uncertainty – Week 10
  - Sampling Distributions, Central Limit Theorem, Confidence Intervals • Chapter 6.4, 7.1
  - HW # 8 posted on Wednesday 11/3; Due by noon Wednesday, 11/10
- Uncertainty – Week 11
  - Central Limit Theorem, Hypothesis Testing • Chapter 7.2
  - HW #9 posted on Wednesday 11/10; Due by noon Wednesday 11/17
- Linear Regression with Inference – Week 12
  - Assumptions of Linear Regression for Inference • Chapter 7.3
  - HW #10 posted by Wednesday 11/17; Due by noon Wednesday 12/01
- Linear Regression with Inference – Thanksgiving Break – Week 13
  - Assumptions of Linear Regression for Inference • Chapter 7.3
  - HW #10 posted by Wednesday 11/17; Due by noon Wednesday 12/01
- Linear Regression with Inference – Week 14
  - Inference about Coefficients and Predictions • Chapter 7.3
  - Review for Final Exam
- Week of 12/6 – 12/10 – Final Exam – date/time TBD