DEPARTMENT: Epidemiology

COURSE NUMBER: TBD

COURSE TITLE: Introduction to Monte Carlo Simulation Using R

CREDIT HOURS: 2

SEMESTER: Fall 2024

CLASS HOURS AND LOCATION: TBD

INSTRUCTOR NAME: Ashley I. Naimi

INSTRUCTOR CONTACT INFORMATION

- **EMAIL**: ashley.naimi@emory.edu
- **SCHOOL ADDRESS OR MAILBOX LOCATION**: CNR 4013
- **OFFICE HOURS**: By Appointment

COURSE DESCRIPTION

This course will focus on how to use experimental principles to appropriately the design and analyze Monte Carlo simulation studies. Simulation studies are an invaluable tool in any analyst's kit. They can facilitate developing a firm understanding of basic and advanced statistical concepts, and provide a flexible means of evaluating whether analytical techniques will work as expected under specific conditions.

Simulation methods are extremely flexible, and can be used to understand and evaluate methodology in a number of different ways.

For example, confidence intervals are commonly used to capture the variation in a parameter estimate of interest, but are notoriously difficult to interpret. Simulation can be used to clarify why this is the case, and how to avoid falling in traps of misinterpretation.

In computing the standard error of a point estimate from a regression model, one may often choose between a robust variance (sandwich) estimator, model-based approaches, or the bootstrap. Simulation can be used to evaluate how well each standard error estimator captures the true sampling variation of the parameter in a specific context, thus guiding the choice.

Measurement error of an exposure of interest is commonly encountered in the empirical sciences, yet researchers will often assume certain simple measurement error models that lead to little to no bias. However, the impact of similar sources of error on covariates included in a regression model (e.g., confounders) is often not considered. Simulation can be used to better understand how common sources of error in a given research project can affect the bias of the treatment effect estimator of interest.

When seeking to construct a simulation study to answer a specific question, several problems need to be considered and controlled for. This course will provide insight into what these problems are, and how to resolve them. Such problems include: choosing an appropriate Monte Carlo sample size to efficiently quantify parameters of interest without unnecessarily slowing down computation; choosing a relevant data generating mechanism using causal inference principles (via, e.g., DAGs) for the underlying research question and using R code to generate variables from this mechanism; and how to efficiently analyze

simulated data and interpret results. The course will conclude with a discussion of when more complex simulation designs are warranted, such as “plasmode” simulations or synthetic simulation (via variational autoencoders or generative adversarial networks).

Course concepts will be illustrated through an extended comparison of two average treatment effect estimators: inverse probability weighting and marginal standardization. After briefly reviewing how these estimators work, we will design a simulation study to evaluate their performance relative to one another. Throughout, we will cover how this specific comparison of two ATE estimators generalize to other questions that might be of interest. This specific example will be used to emphasize the general skills needed to conduct simulation studies in a range of topic areas.

This is an applied course. By the end of the course, students will be able to implement their own Monte Carlo simulation to estimate bias, mean squared error, confidence interval coverage, and other statistics for an estimator of their choice.

PRE-REQUISITES

This course will build on basic and intermediate analytic methods covered in EPI 538, EPI 545, and EPI 550.

Prerequesite skills and concepts include: basic epidemiological measures, confounding, misclassification, selection bias, estimation of epidemiological parameters, issues related to causality, interpretation of basic inferential statistics such as p values and confidence intervals, as well as concepts, methods, and application of key regression concepts related to linear, log-linear and logistic regression.

DIVERSITY, EQUITY, AND INCLUSION CONSIDERATIONS

This course will focus on the theory and application of quantitative and statistical methods to epidemiologic data. Epidemiology is a complex field of study that combines biomedical, physiological, mathematical, social, political, and economic dimensions into a single domain. As a result of this complexity, it is important to understand how epidemiologic knowledge is shaped by and is used to shape social and cultural perspectives on health, well-being, and the optimal organization of human societies.

It is impossible to understand these perspectives without acknowledging the role that early 20th century views on race, ethnicity, sex, gender, and other related socio-political constructs played in shaping quantitative methods that we still use today. For example, the "founding fathers" of statistics (Francis Galton, Karl Pearson, and Ronald Fisher) were also founders of 20th century Eugenics, and they used the new math they derived to characterize many of the egregious and scientifically unjustifiable eugenic acts (forced sterilization, marriage prohibitions, or the supposed moral superiority of "Nordics" or "Aryans") with the patina of "objectivity."

The literature on this topic is expansive, complex, and rapidly growing, and we will not be able to cover many of the problems with how statistical and quantitative methods were and are used inappropriately for iniquitous ends. However, the recommended reading list for the course contains references to key books and papers on this topic which are (some highly) recommended. Additionally, a deep understanding of the connection between data, statistics, and substantive theory can go a long way in both dismantling unsubstantiated claims (both eugenic, and more generally), as well as design studies that can generate a more nuanced understanding of the complexity of health. My hope is that EPI 560 will serve this end.

COURSE LEARNING OBJECTIVES

By the end of this course, you should be able to design, analyze, and interpret results from simple to intermediate level simulation studies using the R programming language. Specifically, you should aim to:

- Understand the distinctions between parametric, plasmode, and synthetic simulations
- Be capable of simulating random variables from a set distribution functions in R
- Be capable of simulating data from linear, log-linear, and logistic regression models
- Use the "balancing intercept" in the context of a logistic regression model for simulation
- Construct simulation functions in R and use for loops and the `apply` family of functions
- Use DAGs to generate code for simulating from simple and complex designs
- Use Monte Carlo Integration to compute true parameter values in simple and complex designs
- Understand Monte Carlo Error, and how to compute performance measures that take it into account
- Generate appropriate summaries (tables and plots) of simulation results in R

ATTENDANCE POLICY

In person attendance in this course is expected.

EVALUATION

Assignments

Each section will be associated with a short section assignment to be completed at home. All assignments are "open-book". Use of internet search engines is encouraged. Use of ChatGPT or other large language model based chat-bots is not prohibited. However, dishonest or misleading use of these (or any) techniques will be considered as Academic dishonesty, which will incur associated penalties (see below).

Grade scale

Students will be graded as Satisfactory (S)/Unsatisfactory (U).

- The basis for the final grade will be determined via section assignments.
- There will be a total of five section assignments (section 1 will not have an assignment).
- Each assignment will consist of two questions.
- Section assignments can be worked on in groups, but must be submitted individually.
- Section assignments will consist of a mix of short answer questions and/or R programming exercises.
- Section assignments will be graded as Satisfactory (S)/Unsatisfactory (U).
- To obtain a Satisfactory grade for the course, the student must obtain an S grade on each section assignment.

COURSE LOGISTICS

CANVAS

- All materials for the course will be hosted on CANVAS. This includes the syllabus, lecture notes, data sets, R programs, section assignments, and readings.

- We will use CANVAS Course Chat, CANVAS Announcements, and email as the primary means of communication in this course.

- You will be asked to use CANVAS to submit assignments by the assigned due date.

In Class Computing

You really should bring your laptop to class. There will be several class opportunities to run code.

R and Posit

Students will be expected to have R and Posit (formerly RStudio) installed and working on their computers. In addition, the following packages should be installed and in working order:

```
"tidyverse", "here", "sandwich", "lmtest", "boot", "ggplot2", "broom",
"rio"
```

Other packages will have to be installed during the course of the semester. Students should be familiar with how to install packages in R from CRAN.

Depending on the analytic scenario, you may have to install a development package from, e.g., GitHub. The best way to do this is to use the `install_github()` function in the `remotes` package (the `remotes` package can be installed from CRAN). However, you will have to address the potential GitHub API limits, which can lead to installation errors. To deal with this problem, you will need your own GitHub account.

The easiest way to address this issue is to use a Github personal access token (PAT). There are a number of ways to do this, and it's important to read the basic information on PATs. Within R and RStudio, one straightforward way to manage PATs is to install and use the `usethis` package, which has a suite of functions available for creating and integrating PATs. Once you've installed `usethis`, you can:

- Use `usethis::browse_github_pat()` to create a GitHub token
- Use `usethis::edit_r_environ()` and add the environment variable by adding the following line to the R environment file: `GITHUB_PAT = 'your_github_token'`.
- Restart R (so that the GITHUB_PAT is read) and try to reinstall the packages that were resulting in the API limit error.

Be aware: **your Github PAT is a password, and should be treated as such.**

Health considerations

At the very first sign of not feeling well, please stay at home. You will not be penalized for not showing up to class because you are feeling ill.

OTHER POLICIES

Accessibility and Accommodations

As the instructor of this course, I endeavor to provide an inclusive learning environment. I want every student to succeed. The Department of Accessibility Services (DAS) works with students who have disabilities to provide reasonable accommodations. It is your responsibility to request accommodations. In order to receive consideration for reasonable accommodations, you must register with the DAS at https://accessibility.emory.edu/students/. Accommodations cannot be retroactively applied so you need to

contact DAS as early as possible and contact us as early as possible in the semester to discuss the plan for implementation of your accommodations. For additional information about accessibility and accommodations, please contact the DAS at (404) 727-9877 or accessibility@emory.edu.

Laney Academic Integrity Statement

You are expected to uphold and cooperate in maintaining academic integrity as a member of the Laney Graduate School. By taking this course, you affirm your commitment to the Laney Graduate School Honor Code, which you can find in the Laney Graduate School Handbook. You should ensure that you are familiar with the rights and responsibilities of members of our academic community and with policies that apply to students as members of our academic community. Any individual, when they suspect that an offense of academic misconduct has occurred, shall report this suspected breach to the appropriate Director of Graduate Studies, Program Director, or Dean of the Laney Graduate School. If an allegation is reported to a Director of Graduate Studies or a Program Director, they are in turn required to report the allegation to the Dean of Laney Graduate School.

COURSE MATERIALS

Required Reading

The course notes will be posted on CANVAS and are the only required readings for this course.

Recommened Optional Reading

1. Rudolph et al (2021) Simulation as a Tool for Teaching and Learning Epidemiologic Methods. *Am J Epidemiol*. 190(5):900-907.
2. Morris et al (2019) Using simulation studies to evaluate statistical methods. *Stat in Med*. 38(11):2074-2102.
3. Maldonado and Greenland (1997) The importance of critically interpreting simulation studies. *Epidemiology* 8(4):453-6
4. Rudolph et al (2021) Simulation in Practice: The Balancing Intercept. *Am J Epidemiol*. 190(8):1696-1698.
5. Fox et al (2022) Illustrating How to Simulate Data From Directed Acyclic Graphs to Understand Epidemiologic Concepts. *Am J Epidemiol*. 191(7):1300-1306.

COURSE OUTLINE: LECTURES

N.B.: Weekly Schedule is Approximate

§ 1	Topics
Week 1:	Why simulate?; An Overview of Simulation Designs; Example Simulation Questions:
	- Example 1: simple regression in an RCT
	- Example 2: IP-weighting versus marginal standardization
	- Example 3: causal mediation analysis

§ 2	Topics
Week 2:	Key Distributions in the R stats and other packages; The Inverse Transformation Method;
Week 3:	Regression Models and Distributions for Simulation Studies; The Balancing Intercept
Week 4:	Constructing User Defined Functions in R; For Loops and the Apply Family of Functions
Week 5:	Seeds in General and in R

§ 3	Topics
Week 6:	The Aims of a simulation study; Defining your data generating mechanism using DAGs
Week 7:	Defining your data generating mechanism using DAGs; Plasmode Simulation
Week 8:	What is your Estimand? Computing the True Estimand Value Using Monte Carlo Integration and the Oracle

§ 4	Topics
Week 9:	Monte Carlo Error and Approximate Performance Measures: Bias, Mean Squared Error, Integrated Measures
Week 10:	Approximate Performance Measures: Efficiency, Confidence Interval Coverage and Length, Type I and II errors, Power, Statistical Considerations

§ 5	Topics
Week 11:	Analyzing and Interpreting Simulation Results, Nested Loop Plots, Zip Plot, Lollipop Plot, Histograms, Density Plots, Scatter Plots

§ 6	Topics
Week 12:	Revisiting Example Simulation Questions:
	- Example 1: simple regression in an RCT
	- Example 2: IP-weighting versus marginal standardization
	- Example 3: causal mediation analysis

§ 7 (optional)	Topics

§ 7 (optional)	Topics
Week TBD:	Computation: Parallel Processing in General and in R
Week TBD:	Working on a Computing Cluster