University of Pennsylvania
Division of Biostatistics
Subject Guide

BSTA 670: Programming and Computation for Biomedical Data Science

Credit points:	**1.0**
Semester:	**Spring 2024**
Time:	**M/W 10:15-11:35am EST**
Location:	**701 Blockley Hall**

Course Instructor: Kristin A. Linn
Assistant Professor of Biostatistics
Email: klinn@pennmedicine.upenn.edu
Office: 220 Blockley Hall
Office hours: Wednesdays 12-1pm; or by appointment
Location: 220 Blockley

TA Zhuoran Ding
Email: dingzh@pennmedicine.upenn.edu
Office hours: Tuesdays 1-2pm; or by appointment
Location: TBD

Pre-requisites: BSTA 620, 621, and 651; or permission of instructor.

Subject Aims: The course will cover programming and computational fundamentals in Python and R. It will concentrate on computational tools that are useful for statistical research and computationally intensive analyses. The goal is for students to develop a knowledge base and skill set that includes a wide range of modern computational tools needed for statistical research and data science. Topics may include, but are not limited to:

1. Reproducible research and programming
2. Algorithms
3. Simulation
4. Computer storage and arithmetic
5. Numerical Integration
6. Optimization

Course Materials: All course materials will be available on Canvas. Canvas is assessable from the Penn library: https://canvas.upenn.edu

Software: A combination of R and Python will be used.

Textbook: None required.

Breaks: There will be **no class** on:
March 4 and 6 (Spring Break)
March 11 and 13 (ENAR conference)
March 20 (Works in Progress day)

Assessment: All assignment materials will be submitted on Canvas. Grades will be based on the following components:

Problem sets: 60% (4 @15% each)
Final project: 40%

Students are encouraged to discuss strategies for solving problem sets, but all submitted code should reflect each student's unique implementation. **Evidence of shared code will be penalized.**

Late Policy: Late assignments will receive a maximum of half credit. An assignment submitted 1 minute after the deadline will be considered late. Assignments more than 3 days late will not be graded and will receive no credit. **If you have a pre-existing commitment or special circumstance (e.g., conference travel, family emergency) please let me know as far in advance as possible so that we can make alternative arrangements for submitting your work.**

Final Project: PhD students will replicate and extend the results of a recently published Monte Carlo stimulation experiment. The final project will include an R package containing simulation code and a report written in .Rmd that fully reproduces the simulation experiment.

MS students will have the option to complete the simulation project described above or perform an applied analysis of a public data set in a Python notebook or Rmarkdown document.

Additional details about the final project requirements will be given later in the semester.

All students will present their final project during one of two in-class presentation days: April 29 and May 1, 2024. **Attendance is required on both days.** 10 points will be deducted from the final project grade for each absence on these two days unless the absence is approved in advance by Dr. Linn.

All project materials due on Canvas: May 3, 2024, by 11:59pm EST

Use of Generative AI Tools

I encourage you to use foundation models such as ChatGPT, GitHub Copilot, etc., in combination with critical thinking skills to further your educational development. If you use these models to obtain quick solutions, you will be missing out on learning opportunities and potentially stifling your own creativity. Keep in mind large language models may produce incorrect statements and fake citations, and code generation models may produce incorrect outputs. **If you use materials produced by foundation models, you must cite them as you would any other reference materials.** It is also important to "show your work" to get full or partial credit, i.e., please **document what prompts you used** to obtain your outputs.

Useful resources:

Git documentation and book by Chacon and Straub: https://git-scm.com/book/en/v2

Python documentation: https://docs.python.org/3/

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. MIT press.

Wickham, H (2015). *Advanced R*. CRC Press.

Matloff, N (2011). *The Art of R Programming*. No Starch Press.

Monahan, J (2011). *Numerical Methods of Statistics* (second edition). Cambridge University Press.

Givens, G.H., & Hoeting, J.A. (2013) *Computational Statistics*. Second edition. Wiley.

Cheney, W, & Kincaid D. (2008) *Numerical Mathematics and Computing*. Sixth edition. Thomson.

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.