BST 270 Reproducible Data Science
Fall 1 2017 – 11:30am -1:00pm Tuesdays/Thursdays
Kresge 201

Instructor:

Heather Mattie
Instructor of Biostatistics
Harvard T.H. Chan School of Public Health
Building 1 Room 421A, 655 Huntington Ave, Boston MA 02115
hemattie@hsph.harvard.edu
(617) 432-5308
Office hour: TBD or by appointment, Building 1 Room 421A

Purpose of this course

Reproducible research has become increasingly important in the biomedical sciences. The bioscience community has recognized reproducibility is a growing challenge in basic, clinical and population sciences. Experimental design, data provenance, analytic methods and tools, and reporting science play a critical role in the biomedical research ecosystem to ensure scientific rigor, robustness and transparency. Statistical and computational methods and tools are fundamental for making scientific results reproducible.

Course Description

The central theme of the course will be to meet these scientific needs of reproducible science through training in reproducible research. The topics covered in this course include the fundamentals of reproducible science, case studies in reproducible research, data provenance, statistical methods for reproducible science, and computational tools for reproducible science. This is a blended course where students are introduced to course content online through videos and reading assignments, and then discuss the content in lecture. Each student will submit a completely reproducible research project and give a short presentation at the end of the course.

Course Objectives

Upon successful completion of this course, you should be able to:

- Describe the fundamentals and importance of reproducible research
- Assess the reproducibility of others' research
- Create a fully reproducible research project
- Develop new methods and tools for reproducible research

Credits: 2.5 credits

Teaching Assistant:

Zack McCaw
PhD Candidate, Biostatistics
zmccaw@g.harvard.edu
Office hour TBD

Course Structure

The course will consist of 6 modules, building up from the fundamentals of reproducible science through case studies, statistical methods for reproducibility and reproducible reporting. Each lecture will be a discussion of the assigned reading and videos. Participation and engagement is expected for each lecture. The last 3 weeks of the course will be devoted to in-class peer critiques and presentations of projects.

Course Materials

Course videos, electronic copies of course readings, rubrics, guidelines, notes/slides, useful website links and data sets will be posted on the EdX course website. Students will need to create an EdX account and register for the course.

Grading, Progress and Performance

This course assumes substantial and informed student participation. At a minimum, being informed requires class attendance and completion of assigned readings and videos. Class attendance and thoughtful participation are important and will be reflected in part in the final grade. Please notify the instructor of an absence before the class. Performance in the course will be evaluated based on the following:

- Project Proposal (10%)
- Class Participation and Peer Critiques (55%)
- Final Project (25%)
- Final Presentation (10%)

Project Proposal (10%)

A proposal for the final project is due in week 3. Project proposals must include the project topic, details of the data used (empirical or simulated) and scientific questions to be answered, a description of the planned statistical analysis, and which language the code will be written in. A document with further details will be distributed in class.

Class Participation and Peer Critiques (55%)

Class participation will consist of discussions based on the course videos and readings. Students are expected to be prepared and engaged in each discussion.

To facilitate a collaborative environment, students will be separated into groups to participate in peer critiques and provide feedback for the projects of their fellow classmates prior to the final presentations (weeks 6 and 7). Rubrics for the critiques and feedback will be provided by the instructor.

Final Project (25%)

The final project must be submitted in the form of a publishable paper and include an abstract, introduction, methods, results, and discussion section. Visuals and schematics should be included if appropriate. In addition, the code used must be submitted in a reproducible format, and all project documents uploaded to a class GitHub repository. Guidelines for the final project will be provided in a separate document.

Final Presentation (10%)

Final project presentations will take place in weeks 7 and 8. Each student will have 10-15 minutes to present their project topic, data, analysis, and reproducible methods to the class. A 5-minute question and answer session will follow each presentation.

Harvard Chan Policies and Expectations

Inclusivity Statement

Diversity and inclusiveness are fundamental to public health education and practice. It is a requirement that you have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Academic Integrity

Each student in this course is expected to abide by the Harvard University and the Harvard. T.H. Chan School of Public Health Codes of Academic Integrity. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the Student Handbook for additional policies related to academic integrity and disciplinary actions.

Accommodations for Students with Disabilities

Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact the OSA **studentaffairs@hsph.harvard.edu** in all cases, including temporary disabilities.

Course Evaluations

Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement.

Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.

Course Schedule Outline

Week	Date	Topics	Assessments
1	August 29	Module 1: Introduction to Course • Overview • Introduction to faculty • Project assignments	Individual introductions In-class activity
	August 31	Module 2: Fundamentals of Reproducible Science • Why reproducible research matters • Definitions and concepts • Factors affecting reproducibility	Case study In-class activity
2	September 5	Module 3: Case Studies in Reproducible Research • Potti 2006 • Baggerly and Coombes • Ioannidis 2009	Case study In-class activity

	September 7	Module 3: Reproducible Reporting • Journals and reproducible research • NIH guidance on reproducible research	Case study In-class activity
3	September 12	Module 4: Data Provenance • Project design • Journal requirements and mechanisms • Repositories • Privacy and security	Case Study In-class activity
	September 14	*Project Abstracts Due Module 5: Statistical Methods for Reproducible Science • Prediction Models • Coefficient of determination • Brier score	Example project code In-class activity
4	September 19	Module 5 Continued • AUC • Concordance in survival analysis	Example project code In-class activity
	September 21	Module 5 Continued • Cross validation • Bootstrap	Example project code In-class activity
5	September 26	Module 6: Computational Tools for Reproducible Science • R and Rstudio • Python • Git and GitHub Creating a repository	Example project code In-class activity
	September 28	Module 6 Continued • Data sources • Dynamic report generation	In-class activity

		• Workflows Course videos conclusion	
6	October 3	Peer Critiques	In-class group activity
	October 5	Peer Critiques	In-class group activity
7	October 10	Peer Critiques	In-class group activity
	October 12	Project Presentations	15-20 minute presentation, 5 minute question and answer session
8	October 17	Project Presentations	15-20 minute presentation, 5 minute question and answer session
	October 19	Project Presentations *Final Projects Due	15-20 minute presentation, 5 minute question and answer session