MCB 293S: Foundations of Biostatistical Practice

Sandrine Dudoit and Kelly Street

Wednesday 18th October, 2017

BASIC COURSE INFORMATION

Course meeting dates and times

The course consists of a 5-week module, with 2 hours of instruction per week: 1 hour lecture and 1 hour discussion/computer lab.

The course is to be taught every spring semester, starting with Spring 2018.

Course location

TBD.

Instructor

The course will be designed and developed by Professor Sandrine Dudoit (Division of Biostatistics and Department of Statistics, UC Berkeley; https://www.stat.berkeley.edu/~sandrine/) and her PhD student Kelly Street.

The instructor for Spring 2018 will be Kelly Street.

Kelly Street is a PhD candidate in the Graduate Group in Biostatistics at UC Berkeley. His research interests concern the development and application of statistical methods and software for the analysis of single-cell transcriptome sequence (RNA-Seq) data, with particular emphasis on the inference of cell lineages. He has extensive teaching experience as a graduate student instructor (GSI) for both undergraduate and graduate statistics courses. He has developed and taught short courses for the Computational Genomics Resource Laboratory's (CGRL) RNA-Seq Data Analysis Workshop. Along with fellow Biostatistics PhD students, he has taken the initiative to develop and teach Special Topics in Applied Biostatistical Practice and Special Topics in Biostatistical Theory (PB HLTH 298).

Instructor availability

Instructor may be contacted by e-mail: kstreet@berkeley.edu.

COURSE DESCRIPTION

Course prerequisites

Being enrolled in the PhD program in Molecular and Cell Biology and having taken MCB 200A-B.

Overview of course

This module is designed to introduce students to the foundations of statistics in the context of biological research. Rather than focusing on a catalog of specific methods (by essence non-exhaustive and rapidly outdated), the module emphasizes general concepts and approaches necessary for sound statistical practice.

Topics covered include: exploratory data analysis (EDA); data visualization; inferential reasoning; models and assumptions; statistical computing; computationally reproducible research. The statistical methods and software are motivated by and illustrated on data structures that arise in current biological and medical research.

Learning objectives

- Perform EDA on a variety of data structures encountered in biological research, for quality control, identifying the main features of the data, and investigating the plausibility of model assumptions.
- Select appropriate visualization methods to effectively display data.
- Be capable of inferential reasoning.
- Translate a biological subject-matter question into a precise statistical question.
- Think critically about models and their underlying assumptions.
- Starting from a subject-matter question and data, identity appropriate statistical analysis methods.
- Acquire familiarity with the R statistical computing language.
- Carry out research in a computationally reproducible manner.

Methods of instruction

Lecture, discussion/computer lab, group discussion, case studies.

There is no required textbook. All course materials (lecture notes, computer labs, references) will be provided on the course website.

Supplemental reading

D. A. Freedman, R. Pisani, and R. A. Purves (2007). *Statistics*. 4th edition. W.W. Norton & Company, New York.

Online course materials from Data 8 – The Foundations of Data Science: `http://data8.org`.

Other references to be posted on course website.

Websites and links

Course website URL TBD.

- Data 8 – The Foundations of Data Science: `http://data8.org`.
- R Project: `https://www.r-project.org`.
- R Studio: `https://www.rstudio.com`.
- Bioconductor Project: `http://www.bioconductor.org`.

- Project Jupyter: `http://jupyter.org`.
- UC Berkeley Statistical Computing Facility: `http://statistics.berkeley.edu/computing`.

GRADING AND EVALUATION PROCEDURES

This is a one-unit course, with ESU grading basis.

Attendance is required to receive credit for the course. Signatures of students will be collected at the end of each class to demonstrate attendance. If you miss one class, you can make it up by writing a 1–2 page commentary on the material covered in that class. If you must miss more than one class, it must be for research-related purposes (e.g., travel to an investigation site or a meeting for presentation). You will still be required to make up the class as described above.

There is one required final project, which involves data analysis using the methods covered in class and the R programming language.

SCHEDULE

Week 1.

- Exploratory data analysis (EDA).
- Visualization.
- Case studies.

Week 2.

- Inferential reasoning.
- Translation of biological subject-matter question into a precise statistical question.
- Case studies.

Week 3.

- Models and assumptions.
- Regression, e.g., ANOVA.
- Density estimation, e.g., maximum likelihood estimation.
- Hypothesis testing.
- Case studies.

Week 4.

- Models and assumptions.
- Regression, e.g., ANOVA.
- Density estimation, e.g., maximum likelihood estimation.
- Hypothesis testing.
- Case studies.

Week 5.

- Computationally reproducible research.
- Statistical computing with R and Bioconductor software.
- Case studies.

EVALUATION OF COURSE AND INSTRUCTOR

It is UC Berkeley policy that all courses be evaluated as part of an overall campus mandate to evaluate and improve the quality of teaching. Evaluation responses are reviewed by the program director after the course ends and after final grades are turned in and filed. The student evaluations are not designed to measure learning, but they do provide feedback in a variety of areas that affect the learning process.

POLICIES

Classroom Decorum:

- No eating (unless allowed in room);
- Turn off cell phones;
- Ground rules for discussion – respect.

The syllabus and schedule are subject to change.