BSTA630: Statistical Methods and Data Analysis I

Fall Semester, 2023

Introduction

This course is an introduction to statistical methods and inference for **Biostatistics degree candidates, with a mix of theory and application**. It covers basic theoretical concepts, estimation, hypothesis testing, one and two-sample tests, analysis of variance (ANOVA), categorical data analysis, linear and logistic regression models, and survival data analysis. Statistical analyses are implemented primarily using R (`https://www.r-project.org`). A companion SAS (`https://www.sas.com/en_us/software/stat.html`) codes are available for most lectures.

Note: A working knowledge of calculus and linear algebra and one introductory statistics course are required. Students are expected to have formal training in these areas, and/or to receive permission from the instructor prior to enrolling. In addition, familiarity with concepts in probability is needed while these concepts will be only reviewed briefly at the beginning of the course. BSTA 630 is a core course in the Biostatistics graduate program, with the objective to prepare the students for the qualify exam and dissertation research. At the end of the course, students are expected to be able to apply the methods **and derive some of the theories**. **Many students report that it requires a lot of work. For non-Biostatistics students, please carefully consider if this fits your goal/need and if the workload is appropriate for you.**

Description

- Classroom: BRB 252
- Lecture time: 10:15-11:45AM Tuesday and Thursday.
- Instructors: Yimei Li (yimeili@pennmedicine.upenn.edu, liy3@chop.edu, Blockley 626); Rui Feng (ruifeng@upenn.edu, Blockley 209); Instructor office hours (virtual): by appointment
- TA: Jiewen Liu (Jiewen.Liu@pennmedicine.upenn.edu); TA office hour: TBD, via Zoom

Textbooks

- Required textbook: Fundamentals of Biostatistics, 8th Edition (Bernard Rosner).
- Not all materials/homework are covered by this textbook.
- Recommended textbook: Statistical Inference, 2nd Edition (Casella & Berger)

Course Goals

- Become familiar with exploratory data analysis, estimation, hypothesis testing, linear models, logistic models, and survival data analysis.
- **Be able to derive important estimators and their theoretical properties.**
- Develop data analytic skills including familiarity with at least one statistical software program.
- Know how to select and evaluate appropriate methods.
- Understand the relationship among methods.
- Be able to interpret statistical results to other professionals.
- Improve and polish writing skills needed to communicate results of data analyses.

Special Instructions

- Statistical methods will be implemented using statistical software R (SAS codes may be available for some lectures).
- Lecture notes, programs and datasets for examples will be available on `https://canvas.upenn.edu`.
- `R` online resources are available at CRAN: `http://www.r-project.org`.
- SAS online documentation is available at `http://support.sas.com/documentation/`.

Evaluation

- Around 10 homework assignments (40%);
- Data analysis project (15%);
- Midterm in-class closed-book exam (20%);
- Final in-class closed-book exam (25%).

Grading Scale

$90- < 92.5$ A-	$92.5- < 97.5$ A	$97.5- \leq 100$ A+
$80- < 82.5$ B-	$82.5- < 87.5$ B	$87.5- < 90$ B+
$70- < 72.5$ C-	$72.5- < 77.5$ C	$77.5- < 80$ C+
$60- < 70$ D	$0-- < 60$ F	

Homework

- Homework problems are from the following two sources: taken directly from the textbook or developed by the instructors.
- To reinforce and review concepts presented in class.
- Some may require an extension or application of the methodological concepts developed in class or in the text.
- Some may be based on scenarios encountered by collaborative statisticians and require creative thinking and integration of concepts to solve. At least once during the course we will ask you to construct a simulation to evaluate a statistical approach.
- Submit homework through Canvas. Homework is due one week after it is assigned or announced by the instructor otherwise. Late homework will NOT be accepted.

Data Analysis Project

Each student is required to work on one real data project provided by the instructor. Over the course of the semester the students will determine which techniques introduced in the course are appropriate for describing the data and answering the scientific questions of interest. All students will be required to submit a written report at the end of the semester.

Exam

- In class and close-book
- Certain distribution and formulas provided
- Two sheets of cheat sheet allowed

- Suggest bring a calculator; but no electronic device allowed
- Midterm, two hours
- Final, three hours

Tentative Schedule -2023

Date	Lecture	Topic
Aug 29	L1	Probability
Aug 31	L2	Random variables
Sep 5	L3	Probability distribution
Sep 7	L4	Point estimate
Sep 12	L5	Summarizing data
Sep 14	L6	CLT; MOM
Sep 19	L7	MLE
Sep 21	L8	Evaluate estimators; Bayesian
Sep 26	L9	Interval estimation
Sep 28	L10	Hypothesis testing - one sample
Oct 3	L11	Hypothesis testing - two sample
Oct 5	L12	Nonparametric
Oct 10	Review	Review
Oct 12	No class	Fall break
Oct 17	Midterm	In class exam 10:15-12:15
Oct 19	L13	LRT
Oct 24	L14	One-way ANOVA
Oct 26	L15	Two-way ANOVA
Oct 31	L16	Simple linear regression
Nov 2	L17	Regression coefficient and prediction
Nov 7	L18	Multiple linear regression
Nov 9	L19	Goodness of fit
Nov 14	L20	Categorical - part 1
Nov 16	L21	Categorical - part 2
Nov 21	L22	Logistic regression
Nov 23	No class	Thanksgiving break
Nov 28	L23	Sample size and power
Nov 30	L24	Survival data analysis?
Dec 5	Review	Review
Dec 7	Final	In class exam 9-12?