PUBHLTH 490ST (3 credits)
Telling Stories with Data: Statistics, Modeling, and Data Visualization
Fall 2016 :: T/Th 10:00-11:15am :: LGRT 204

INSTRUCTOR
Nicholas G Reich
425 Arnold House
(413) 545-4534
nick [at] umass.edu
Reich Lab website
on twitter: @reichlab

Office Hours: TBD
Teaching assistant: Zhenning Kang
Course website
Piazza site

MATERIALS

Required Textbook

Kaplan, Daniel T. 2011. *Statistical Modeling: A Fresh Approach, 2nd Ed.*.

Recommended Textbook (freely available online)

Diez D, Barr C, and Çetinkaya-Rundel M. 2015. *OpenIntro Statistics, 3rd Ed.*.

Software

R :: r-project.org (or just Google "r")
RStudio :: rstudio.org

Prerequisites

One of any of the following introductory stats courses taught at UMass: BIOSTAT 391B, STAT 111, STAT 240, STAT 501, ResEcon 212, PSYCH 240. If you have not taken an intro stats course at UMass but still want to enroll in this course, you are encouraged to petition the instructor for permission, especially if any of the following apply: (a) you have taken AP Stats in high school, (b) you have taken a college-level intro stats course just not one of the ones listed above, or (c) you are confident in your quantitative skills and your ability to succeed in a fast-paced, advanced introductory course. Additionally, prior programming experience with R or concurrent enrollment in PUBHLTH 497D (Introduction to Statistical Computing with R) is required.

Course Description

The aim of this course is to provide students with the skills necessary to tell interesting and useful stories in real-world encounters with data. Specifically, they will develop the statistical and programming expertise necessary to analyze datasets with complex relationships between variables. Students will gain hands-on experience summarizing, visualizing, modeling, and analyzing data. Students will learn how to build statistical models that can be used to describe and evaluate multidimensional relationships that exist in the real world. Specific methods covered will include linear and logistic regression. Students will work with the R statistical computing language and by the end of the course will require substantial independent programming. The course will not provide explicit or detailed training in R programming. To the extent possible, the course will draw on real datasets from biological and biomedical applications. This course is designed for students who are looking for a second course in applied statistics/biostatistics (e.g. beyond BIOSTATS 391B or STAT 240), or an accelerated introduction to statistics and modern statistical computing.

Learning Goals *(By the end of the course students will be able to...)*

- design data-driven experiments to answer specific questions,
- use data to identify and distinguish patterns of randomness vs. non-randomness,
- create powerful data visualizations that reveal and highlight important features of data or models,
- understand and critique statistical model equations as representations of a given real-world setting,
- formulate, fit, and interpret statistical models to designed to answer specific scientific questions,
- weigh evidence for/against hypotheses about associations between variables,

- diagnose the appropriateness or "goodness-of-fit" of a given model,
- write concise, professional, and reproducible statistical analysis reports.

Expectations

This course will require you to work thoughtfully, carefully, and independently and will require substantial work outside of class time. Because we will be using a more project-driven approach in this course, with assignments that will build upon one another into a final product, it is vital that you do not fall behind. If you feel as though you are falling behind or starting to lose a handle on the content, I expect you to come talk to me either after class or during office hours so that I can help as much as I can to set you back on track. Please do not wait to talk to me if you start to fall behind.

I also expect you to devote substantial outside-of-class time to your work for this course, typically involving 5-10 hours per week. I anticipate that this work will be divided among:

- finishing in-class activities
- reading assigned articles and chapters
- reviewing your notes
- working on assignments
- conducting project work
- preparing for exams

Things you should expect from me:

- timely feedback on assignments and quizzes
- response to questions via Piazza or email in < 2 working days (often sooner)
- attention to your questions related to coursework during office hours
- instruction in how to write, research, and debug R code

Things you should not expect from me:

- time for frequent non-office hour drop-in questions
- comments on a research project that is unrelated to your coursework
- writing your code for you or *extensive* debugging of your code

Types of Assignments and Activities, with Grade Contributions

Homework (40%): There will be approximately six lab assignments that you will complete over the course of the semester. Each lab will have components that you will hand in for grading. Assignments and due dates will be posted in advance on the course website. The assignments will be graded. Some assignments will require you to submit a digital file with reproducible solutions, i.e. a knitr file that reproduces your answers. Late homeworks will not be accepted under any circumstances. If a homework is not handed in on time, it will receive a grade of zero. I will drop your lowest homework grade when calculating your final grade.

Midterm exam (25%): There will be an in-class mid-term exam in this course. You will be allowed one single-sided, 8.5x11, sheet of notes for the exam.

Final Project (25%): In the second half of the course, you will develop and write your own data story. This project will be presented to your classmates. A separate handout will provide details.

Participation/citizenship (10%) : Being a good class "citizen" plays a large role in your final grade. A few of the characteristics of good class citizens are: attending all course meetings, using office hours, asking questions, offering to answer questions, actively listening when others are talking, and participating on Piazza (both asking and answering questions). Citizenship is more a function of quality than quantity. The "default" citizenship score is 5 out of 10.[1]

Extra Credit: If you find a mistake in the course materials or make an improvement (as judged by the instructor), and submit the update as a pull request via GitHub, you will receive one point of extra credit on your final grade per distinct accepted pull request (up to a limit of 5 pull-request extra points). If you send me an email with "I read the syllabus" as the subject line by the beginning of the second class, you will receive two points of extra credit on your final grade.

[1] Acknowledgments to Aaron Swoboda for introducing me to the concept of course citizenship and for some of this text.

Course Policies

Collaboration on all assignments is expected and encouraged, although you must write up your own assignment. **No copying or cutting and pasting.** Each project will contain an independent component which must be your own work. You may discuss your project with others and even solicit ideas and advice, but at the end of the day, you must complete all the analysis and write-up on your own. Any explicitly borrowed ideas or language must be cited or acknowledged appropriately.

Attendance is required. Absences (excused or not) will impact your participation grade.

All mobile devices that can/will be distracting to you or others during class must be turned off at the start of class and may not be used during class time.

Course Schedule

This is a tentative course schedule and is subject to change with little or no notice.

Unit 1 What is data?
Introduction, motivation, and overview
Design of experiment, data collection
Unit 2 Exploratory Data Analysis
Summarizing and visualizing data
Unit 3 Introduction to Models
The Language of Models
Unit 4 Linear regression
Simple linear regression
Multiple linear regression
Midterm review and exam
Unit 5 Weighing evidence from models
Confidence and uncertainty in models
Hypothesis testing
Unit 6 Modeling binary outcomes
Logistic Regression
Final projects

Grading Scale

Grade	Percentage
A	93-100
A-	90-92
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	0-59

Council on Education for Public Health (CEPH) Course Competencies

- Distinguish among the different measurement scales and the implications for selection of statistical methods to be used based on these distinctions.
- Describe conceptual frameworks (statistical literacy) in biostatistics
- Apply biostatistical methods to the design of studies in public health.
- Use computers to appropriately store, manage, manipulate and process data for a research study using modern software.
- Apply descriptive techniques commonly used to summarize public health data.
- Describe the basic concepts of probability, random variation and selected, commonly used, probability distributions.
- Select and perform the appropriate descriptive and inferential statistical methods in selected basic study design settings.
- Describe appropriate methodological alternatives to commonly used statistical methods when assumptions are violated.
- Integrate analysis strategies in biostatistics with principles and issues in epidemiology. literature
- Develop written and oral presentations based on statistical analyses for both public health professionals and educated lay audiences.
- Apply statistical methods to solve problems in the health sciences and carry out theoretical research in statistical methodology.

Academic Honesty Policy Statement

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst. Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. The procedures outlined below are intended to provide an efficient and orderly process by which action may be taken if it appears that academic dishonesty has occurred and by which students may appeal such actions. Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent. For more information about what constitutes academic dishonesty, please see the Dean of Students' website.

Disability Statement

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier-free campus. If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. It is your responsibility to initiate these services and to communicate with faculty ahead of time to manage accommodations in a timely manner. For more information, consult the Disability Services website.