Introduction to Data Analysis
Syllabus

Garrett Darl Lewis

December 3, 2019

Course Overview

This course will provide you with an introduction to the burgeoning field of data analysis and related research methods. We will cover approaches and techniques for obtaining, organizing, exploring, and analyzing data using elements of statistics, machine learning, and statistical computing. While these tools are applicable across many different fields, this course will focus primarily on applications to the social sciences.

Instructor

Darl Lewis
Email: glewis@princeton.edu
Office: 224 Robertson
Office Hours: TR 3:00 PM – 5:00 PM or by appointment

Lectures: MWF 11:00 AM – 11:50 AM, Lecture Hall
Website: https://my.course.website/

Materials and Prerequisites

There are no official prerequisites for this course. However, it is recommended that students be comfortable with the basics of command-line computer usage and undergraduate-level mathematical notation and symbols (e.g., summation and product symbols, algebraic notation).

There are no official textbooks for this course, however, I will post optional readings from the following:

- Jim Pitman. 1993. Probability. Springer, New York, Berlin, and Heidelberg. ISBN: 0-387-97974-8.
- Richard J. Larsen and Morris L. Marx. 2012. An Introduction to Mathematical Statistics and Its Applications, fifth edition. Prentice Hall. ISBN: 0-321-69394-9.

Lastly, you will need a fully functional and up-to-date computer. I do not recommend attempting the computational assignments for this class with a phone, tablet, or other similar device. You do not need the most advanced specifications available, but it should be a device that is capable of running a standard suite of modern software.

Assignments and Grading

Your grade will be determined as follows:

- Three problem sets worth 10% each, due at 10:59 AM on the 2nd, 4th, and 6th Fridays of the semester
- Two projects worth 20% each, due at 10:59 AM on the 9th Friday and the first day of finals
- Midterm worth 20% on the 6th Friday of the semester
- Final quiz worth 10% on the final day of class*

**In order to incentivize attendance, the open-book final quiz will cover topics addressed in lecture, but not necessarily in the readings. It should be relatively easy in you have regularly attended class, but prove more difficult otherwise.*

Table 1: Minimum Letter Grades

A+	$\geq 97\%$	A	$\geq 93\%$	A-	$\geq 90\%$
B+	$\geq 87\%$	B	$\geq 83\%$	B-	$\geq 80\%$
C+	$\geq 77\%$	C	$\geq 73\%$	C-	$\geq 70\%$
D+	$\geq 67\%$	D	$\geq 63\%$	D-	$\geq 60\%$
		F	$< 60\%$		

Letter grades will follow the standard rubric as listed in Table 1. The listed scores reflect the minimum letter grade that will be assigned for a given cumulative score. However, if necessary (as is likely) a curve will be applied to the final scores. This curve will only be applied to final scores at the end of the semester.

Assignments should be turned in to the lock box located in the departmental office during regular business hours or turned in to me prior to the lecture at which they are due. To help you keep track of your progress, I will return grades assignments no later that the first lecture occurring at least one week after the assignment due date. Any requests for regrades must be made within one week of this time (the first lecture occurring at least two weeks after the due date).

Note: If you find yourself struggling, I encourage you to contact me as soon as possible, especially with respect to the programming aspects of the class, since they are cumulative and it will become quite difficult to catch up if you fall behind.

Problem Sets

Problem sets will generally consist of several standalone problems with closed form solutions. They should be typed and should include any relevant code used in your solutions. Sets should be clearly organized and include all appropriate identification. Points may be deducted for disorganized write-ups or those that fail to show your work. For these assignments, you may collaborate with other students, but you must write up your solutions individually.

Projects

Projects will be somewhat longer than problem sets and leave more room for open-ended responses. However, you are also permitted to work in groups of up to three people if you so desire. Formatting should be similar to problem sets.

Quizzes and Exams

Quizzes and exams may be handwritten, but must nonetheless be organized and legible. Unless otherwise specified, these exams will be open-book and open-notes, however no internet or personal communication will be allowed.

Late Work

As a rule, late work is not accepted. At the discretion of the Head TA, late homework turned in the day it is due, but after the 4:00 pm deadline will be accepted with a 25% penalty. If there are extenuating circumstances, you must obtain approval from the instructor before midnight the night before it is due and you must get a note from the from the Dean supporting the extension.

Schedule

This course will be divided into two concurrent parts. The first will focus on analytic statistics and probability, while the second will focus on statistical computing. Typically, I will focus on the former during the Monday and Wednesday lectures and the latter on Friday, however there will be a significant amount of overlap which will become more pronounced as the semester progresses.

Prior to each lecture, I will post a draft of the relevant notes. This draft will often be intentionally incomplete but will be updated shortly after the lecture. To ensure you have the most up-to-date version, check the time stamp at the top of each document.

Probability and Statistics

Probability will be covered in the first half of the term and statistics in the second half (see below for information regarding textbooks). Specific topics are listed in Table 2. Besides the textbooks listed above, I will also post readings from several alternative sources on the course website. These readings will be announced as they are posted.

Table 2: Schedule of Topics

Week	Topic	Details
1 2	Properties of probability	Random variables, distributions, densities, and expectation Independence, conditional probability, Bayes' Law Joint distributions, marginals, covariance, correlation Summary statistics
3	Important distributions	Bernoulli, Binomial Uniform Normal (Gaussian) Exponential, Poisson Gamma, Beta, Chi-square
5 6	Estimation of parameters	The Law of Large Numbers The Central Limit Theorem Consistency, unbiasedness Maximum likelihood estimation Confidence intervals
7 8	Significance tests	Likelihood ratio tests Monotone Likelihood Ratio Property and the Neyman-Person Lemma Type I and Type II errors Power and assurance Critical values
9	Specification tests	Kolmogorov-Smirnov χ^2 test, Fisher's exact test
10 11	Linear regression analysis	Gauss Markov-Theorem ANOVA
12	Bayesian approaches	TBD

Statistical Computing

Although you are permitted to use any statistical computing software you desire, I will teach this course using the R programming language and offer support for this language. This is an open source language developed specifically for statistical analysis. I will go over the use of this language in Friday lectures.

Supplementary Materials

This course only provides an introduction to statistical analysis. For those looking for more on the material we cover or different angles on the same material, here are some of my recommendations:

- Robert V. Hogg, Elliot A. Tanis, and Dale Zimmerman. 2015. Probability and Statistical Inference. Pearson, Boston. ISBN: 978-0-321-92327-1.
- Alex Reinhart. 2015. Statistics Done Wrong: The Woefully Complete Guide. No Starch Press, San Francisco. ISBN: 978-1-59327-620-1.
- Calvin Dytham. 2011. Choosing and Using Statistics: A Biologist's Guide. Wiley-Blackwell. ISBN: 978-1-4051-9839-4.
- David E. Matthews and Vernojn T. Farewell. 2015. Using and Understanding Medical Statistics. Karger, Basel. ISBN: 978-3-318-05458-3.
- Robert B. Ash. 2008. Basic Probability Theory. Dover, Mineola, New York. Reprint of the 1970 edition published by John Wiley and Sons. ISBN: 0-486-46628-0.
- John B. Walsh. 2012. Knowing the Odds: An Introduction to Probability. American Mathematical Society, Providence, Rhode Island. ISBN: 978-0-8218-8532-1.
- Kai Lai Chung and Farid Ait-Sahlia. 2003. Elementary Probability Theory with Stochastic Processes and an Introduction to Mathematical Finance. Springer-Verlag, New York, Heidelberg, and Berlin. ISBN: 978-0-387-95578-0.
- Richard Isaac. 1995. The Pleasures of Probability. Springer-Verlag, New York, Berlin, and Heidelberg. ISBN: 0-387-94415-X.

Modern statistics is computationally intensive, so you will have to use computers to do some of the assignments. There are a wide variety of languages and platforms available for this. I will be using and teaching R for this course; however, I will not explicitly require you to use this language. Note that if you do decide to use a difference language, I may be limited in the amount of support that I am able to offer. The following are several highly recommended books on R:

- Claus Thorn Ekstrom. 2011. R Primer. Chapman & Hall/CRC Press.
- Paul Teetor. 2011. R Cookbook. O'Reilly Media. ISBN: 978-0-596-80915-7.
- Joseph Adler. 2012. R in a Nutshell, 2nd edition. O'Reilly Media. ISBN: 978-1449312084.

- Alain F. Zuur, Elena N. Ieno, and Erik H. W. G. Meesters. 2009. A Beginner's Guide to R. Springer Science+Business Media, New York. ISBN: 978-0-387-93836-3.
- Peter Dalgaard. 2008. Introductory Statistics with R, second edition. Springer Science+Business Media, New York. ISBN: 978-0-387-79053-4