Syllabus

General Info

Course MAT217 Applied Statistics
Instructor Charilaos Skiadas (skiadas at hanover dot edu)
Term Winter 2018-2019
Office SCH 111 / LYN 108
Office Hours MWF 2:40pm-3:50pm in LYN 108, and by appointment.
Book OpenIntro Statistics[1] by openintro.org, 3rd edition. Free PDF download from the site. Can also buy from online retailers for around $10.
Website for notes[2], for quizzes[3].
Class times MWF 10:40am-11:50am in LYN120B. Labs on Fridays in CFA112C.

Course Description

Applied Statistics focuses on the statistical study of data, its collection, description and inference based on sample data. Here are some of the kinds of questions we might ask:

- Suppose we would like to predict the winner of some future elections. As we cannot ask every single citizen what they would vote, we would need to rely on taking a smaller sample.
  - How big should this sample be?
  - How should we go about selecting who would be part of the sample?
  - As there is a certain degree of variability in the answers we get (different samples would have given us different answers) what can the sample's results tell us about the possible election results? With what certainty can we predict a winner based on the given sample?
- Suppose 10 people measured the length of the same room, and they all found somewhat different results. What does this tell us about the actual length of the room?
- More broadly, repeating runs of an experiment would produce slightly different values for the same physical quantity (e.g. the speed of light). What does that tell us about the actual value of that quantity?
- Can we identify a literary work as belonging to a particular author based on how often common words appear in it, compared to the author's other works?
- A person claims that they can detect, when drinking tea with milk, whether the milk was poured first in the cup or whether it was the tea. How would we go about testing that person? At what point would we feel reasonably certain that they can or cannot detect it?

[1] https://www.openintro.org/stat/
[2] http://skiadas.github.io/AppliedStatsCourse/site/
[3] https://moodle.hanover.edu/course/view.php?id=648

In this course, we will develop the tools to be able to answer these kinds of questions. Simply put:

> In Statistics our goal is to understand the uncertainty and variability inherent in every experiment/phenomenon/measurement, and to attempt to control that variability.

Broadly speaking, the coarse is divided into three parts. We will start with Descriptive Statistics, which deals with the various ways of presenting data, their summaries and inter-relationships, and the problems one might encounter when doing that, both from using bad graphing techniques and from relying too much on numerical summaries. You will be able to understand the pitfalls when people and the media quote average numbers and percentages, and you will be able to put those numbers into a proper perspective. You will familiarize yourselves with the various types of graphs, their strengths and weaknesses, as well as common steps to make the information from the graphs more clearly presented.

The second, brief, part of the course deals with the design of experiments, and sampling methods. It is an introduction to the methods used to collect data, and the problems that arise. As an example, during the great depression a popular magazine made an extremely wrong prediction about who would win the presidential elections, which led to the downfall of that magazine. Their description was based on a massive survey that ended up getting answers from almost 2 million people, so it seems very surprising that they would get things wrong. We will investigate the mistakes that they made, and why having this enormous sample size didn't necessarily help them. We will also touch briefly on some of the fundamental principles employed in designing a study or experiment.

The final part of the course deals with Inferential Statistics. Inferential Statistics concerns itself with making predictions about a population based on information from a small sample. For instance, when CNN reports that Obama and McCain both have 47% support, based on a sample of 1000 people, what does that really tell us about the voting preference of all Americans? And what is it that makes us certain that those 1000 people that were polled are sufficient to make a prediction? Would we have been able to make a better prediction with more people, or does it not matter how many we have after a while? Can we provide some range of values that we can be pretty sure the candidate's actual percentage would be in? If a baseball player has a better on-base percentage than another player on a particular season, does this mean that they are truly better at getting to base? Or did they just happen to have a better season? At which point is it true skill and not just 'luck'?

In order to understand the mechanics behind Inferential Statistics, we will need to spend some time studying the basics of Probability Theory and the notion of a random variable. Probability Theory is the mathematical study of random phenomena and processes, and it will provide us a tool to deal with a wide range of situations, from sampling from a large population to simply a basketball player shooting from the free-throw line, or even the various tests for diseases and how reliable they are.

Goals

The course has the following main objectives:

- You will learn how to critically think about, analyse and evaluate data, and how to formulate your conclusions.
- You will be using computer technology to analyse real data from various disciplines, often involving large data sets that would be very difficult to analyse in other ways.
- You will work in small groups to complete a term project involving the analysis and presentation of certain data. You will have the chance to present both an oral and a written report at the end of the semester. You will thus demostrate skills in developing a thesis statement, supporting that thesis with logical rationale and quantitative evidence, and presenting that thesis in a convincing fashion, orally and in writing.
- You will be introduced to probability theory, which provides the solid foundations on which all statistical inference procedures are based. This will provide you with an understanding of the nature of symbolic language, formal reasoning, and the process of solving problems by means of abstract modeling.

Course Components

Reading Notes and Practice Problems

On the website you will find a schedule[4] with links to documents for each class day. In those documents you will find notes for the day's lesson, a reading assignment, and a list of practice problems. You should work on those practice problems, and ask any questions you have about them. You do not have to turn the problems in.

Class Attendance

You are expected to attend every class meeting, including labs. You are only allowed to miss 3 classes without excuse. From that point on, every unexcused absence will result in a reduction of your final score by one percentage point, up to a total of 5 points. Excused absences should be arranged in advance, and backed by appropriate documentation. Emergencies will be dealt with on an individual basis. There are very few reasons that would qualify as an excuse for an absence.

Homework Assignments

Around once or twice a week, I will be assigning homework. These will be collected, and counted on a completion scale of $0, 0.25, 0.5, 0.75, 1$, depending on how much effort you have put and how complete your work is. Questions on the quizzes and exams tend to be similar to the homework problems, so it is to your advantage to really

[4] http://skiadas.github.io/AppliedStatsCourse/site/schedule.html

understand the homework, and not merely "do it" or copy it just to get it turned in. Homework assignments are 5% of your final grade.

Online quizzes

We will be using the Moodle platform[5] for online quizzes. You will typically have one quiz each week. Each quiz has a two hour time limit, and will have a deadline no more than a week after we cover that topic. You are allowed to take the quiz up to 2 times before that deadline, and you receive feedback after each attempt. The average of the two tries will be your final quiz score. You are expected to work on the quizzes on your own, and you are allowed to refer to the book and class notes while taking them. Your quiz score is 10% of your final grade.

Exams

There will be two midterms, on Wednesday, October 4th and Friday, November 10th, and a final/3rd midterm during finals week. **You have to be here for the exams**. If you have conflicts with these days, let me know as soon as possible. Do not plan your vacation before you are aware of the finals schedule. In terms of your final grade, the exams you did better on will weigh more.

Term Project

Throughout the semester, you will work in groups of three on a term project.

- In the first 2 days you will need to pick group-mates and choose from a short list of pre-selected projects.
- At the beginning of the term you will start by creating a series of questions to ask about the given data.
- As we move along the term, you will be applying the skills we learn to analyze the given data and answer key questions about the dataset.
- At the end of the term you will a report and present your conclusions to the other students.

Getting Help

- The learning center has some awesome tutors that can really help you out with the material. USE THEM! But make sure to go prepared to those meetings; you should have specific questions ready. And keep in mind that tutors are there to help you understand the concepts, they are not there to do your homework for you.

[5] http://moodle.hanover.edu

- You should never hesitate to ask me questions. I will never think any less of anyone for asking a question. Stop by my office hours or just email me your question, which has the great benefit of forcing you to write it down in clear terms, which often helps you understand it better.
- You are allowed, and in fact encouraged, to work together and help each other regarding the notes and the practice problems. However, I strongly encourage you to try the problems out on your own first before talking to someone about them.
- You may discuss homework problems with others, but only after you have spent some time trying them on your own. And in any event the submitted work must be your own! So even though you may talk to others about the problem, when you sit down to write the answers you should be on your own.
- Your work on the online quizzes must be your own. You may ask me or the tutors any questions you have related to them, but your final answers must be your own.

Grading

Your final grade depends on class attendance, homework, project, quizzes, midterms and the final, as follows:

Component	Percent
Attendance	5%
Homework	5%
Quizzes	10%
Project	15%
Worst Midterm	15%
Middle Midterm	20%
Best Midterm	30%

This gives a number up to 100, which is then converted to a letter grade based roughly on the following correspondence:

Letter grade	Percentage Range
A, A-	90%-100%
B+, B, B-	80%-90%
C+, C, C-	70%-80%
D+, D, D-	60%-70%
F	0%-60%