Syllabus

General Info

Course MAT217 Applied Statistics
Instructor Charilaos Skiadas (skiadas at hanover dot edu)
Term Winter 2014-2015
Office SCH 121C
Office Hours MWF 3pm-3:30pm, WRF 9am-10am, T 10am-12am
Book Introductory Statistics[1] by Openstax College
Websites for notes[2], for assignments[3]
Class times MWRF 1pm-2pm SCC112. On Wednesdays in CFA112C

Course Description

Applied Statistics focuses on the statistical study of data, its collection, description and inference based on sample data. For instance, suppose presidential election polls these days show the two candidates getting the same percentage of support. What does that really mean? That the candidates would get exactly that percentage, if we were to hold elections this moment? How was this percentage obtained, and how reliable is it? What range should we expect their true percentage to be in (in other words, what is our margin of error)? Suppose 10 people measured the length of the same room, and they all found somewhat different results. What does this tell us about the length of the room? Can we identify a literary work as belonging to a particular author based on how often common words appear in it? How do we determine the speed of light? It is reasonable to believe that different attempts at measuring that speed would provide us with slightly different values, which one is "correct"? 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 and phenomenon, 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 yourself with the various types of graphs, their

[1] http://openstaxcollege.org/textbooks/introductory-statistics/get
[2] skiadas.github.io/AppliedStatsCourse/site/
[3] https://moodle.hanover.edu/course/view.php?id=228

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 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. This will enable you to "make critical reflective judgements".
- You will be using computer technology to analyse real data from various disciplines, often involving large data sets that would be impossible to analyse in other ways.
- You will work in small groups to complete a term project, which will contain the formulation of a research question and methodology, as well as the collection and analysis of the resulting data. You will have the chance to present both

an oral and a written report at the end of the semester. This will enable you to "demonstrate skills in independent thinking by developing your own thesis statement, supporting that thesis with logical rationale and appropriate evidence, and presenting the 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 enable you to "understand the nature of symbolic language, formal reasoning, and the process of solving problems by means of abstract modeling".
- By the study and understanding of random variables, and their ubiquity in modern scientific methodology, you will be able to "identify the essential qualities of these formal tools that underlie their effectiveness in the solution of real-world problems".
- By examining the various assumptions needed for formal inference procedures, and the extent to which they are violated in practice, you will be able to "explain the limitations of these formal systems of reasoning".

Course Components

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 a 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 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) or (1), depending on how much effort you have put and how complete your work is. Questions on the exams tend to be similar to the homework problems, so it is to your advantage to really *understand* the homework, and not merely "do it" or copy it just to get it turned in. Homework assignments are (10%) of your final grade.

Online quizzes

We will be using the Moodle platform at http://moodle.hanover.edu to provide structure for the course. There you will find sections for each topic we cover, containing links to reading material from the book, practice problems from the book, as well as a quiz to test your understanding of the material. Each quiz is untimed, and will have a deadline no more than a week after we cover that topic. You are allowed to take the

quiz up to 3 times before that deadline. Only the best of those 3 tries will count. You are expected to work on the quizzes on your own, and you are allowed to refer to the book or class notes while taking them. Your quiz score is (10%) of your final grade.

Exams

There will be two midterms, on Wednesday, February 5th and Friday, March 14th, 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 at will weigh more.

Term Project

Throughout the semester, you will work in groups of three on a term project. The project consists of four phases:

1. Getting a group together and formulating your research question and methodology. This should happen within the first 2-3 weeks.
2. Collecting the data necessary for answering the question. This should take the next 1-2 weeks.
3. Analysing the data. This should take the next 4-5 weeks.
4. Writing a report and presenting your conclusions to the other students. This will happen during the last 2-3 weeks.

Here is a more specific schedule:

1. In the first week you should find two more people to be your teammates. The teams should email me by the end of the **first week**. Any people not in an assigned team by Monday of the second week will be assigned a team by me. When forming your teams, you will need to also have, and email to me along with the team member names, at least **5 hours** during the week that you can all meet to work on the project, or if that is not possible, then exactly how you are planning to meet and work together on the project. By joining a team you are making a commitment to your team-mates, to work with them over the semester. I expect you to honor that commitment.
2. All teams will present their proposals in class by the end of 2nd week. These proposals should fit it at most 2 pages, plus a third page with the survey questions, prepared in powerpoint, and contain the following:
    - A **project description**. What is the main focus of your project? What kind of relation/phenomenon are you trying to examine with this project? Be **original**.

- A list of the **variables** you will be measuring. You should have, at the very least, two categorical and two quantitative variables, and overall about 6 to 8 variables. These should be accompanied with short list of what interactions you expect between these variables. These expectations can then be used when analyzing the data, as a starting point.
- A description of the methodology you are going to follow for collecting your data. Is it going to be a survey? How will it be administered? Are you going to collect data from the internet? How are you planning to assert the reliability of that data?

I will give you a sample proposal by the end of the first week, so that you have a model to base it on.

3. This proposal should be finalized by the end of the third week, based on feedback from the in-class presentation. In the next two weeks, your data should be collected and assembled in a datasheet. This sheet should be emailed to me no later than the end of the fifth week.

4. From this point on, you should gather together and analyse your data, discussing how it helps you answer the questions, and any pitfalls and problems you might have. **It is very important that this part of the analysis be done with all 3 members working together**.

5. A **final paper report** will be due during the last week. You may send me draft reports before that, if you want to receive feedback on them, and I strongly encourage you to do that.

6. The last two days of classes will be devoted to in-class **presentations** of your project. This is an opportunity to talk about your work to members of the other teams, as well as hear about their projects. **Each** member of the team will have to give talk for at least 5 minutes.

Grading

Your final grade depends on class attendance, labs, project, quizzes, midterms and the final, as follows (the weights on the 3 exams are based on which exams you did better on: Your best exam will count for 30%, your second best for 25% and so on):

Class Attendance	5%
Quizzes	10%
HW	10%
Project	15%
Worst Exam	10%
Middle Exam	20%
Best Exam	30%
Total	100%

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%

\end{document}