Lecture 1: Introduction

- Welcome
- Why is this course important?
- Why is this course useful?
- Logistics
- Probability v.s. Statistics

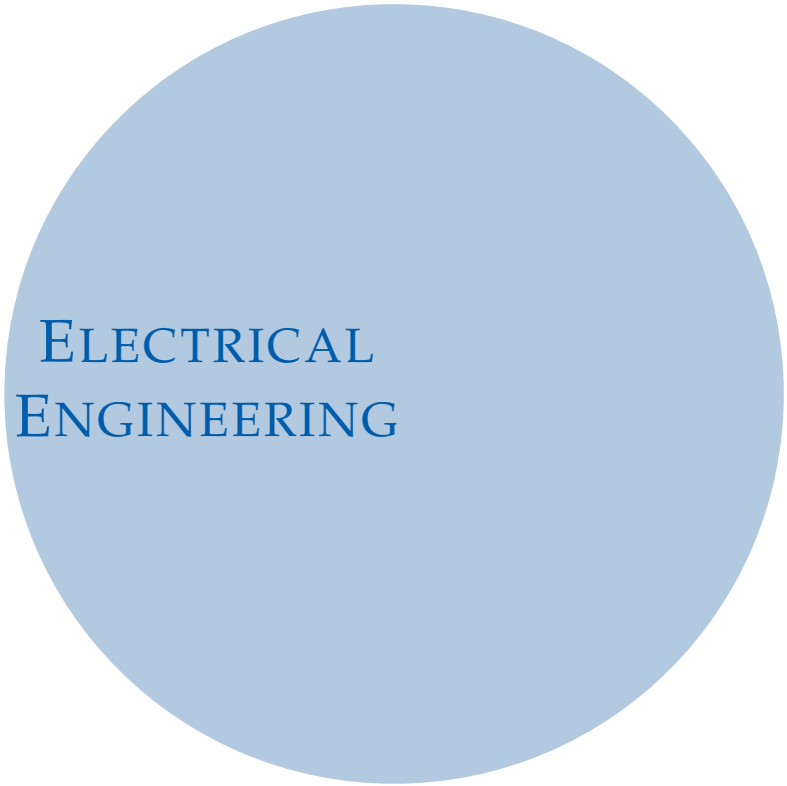
Welcome to STA111

Introduction

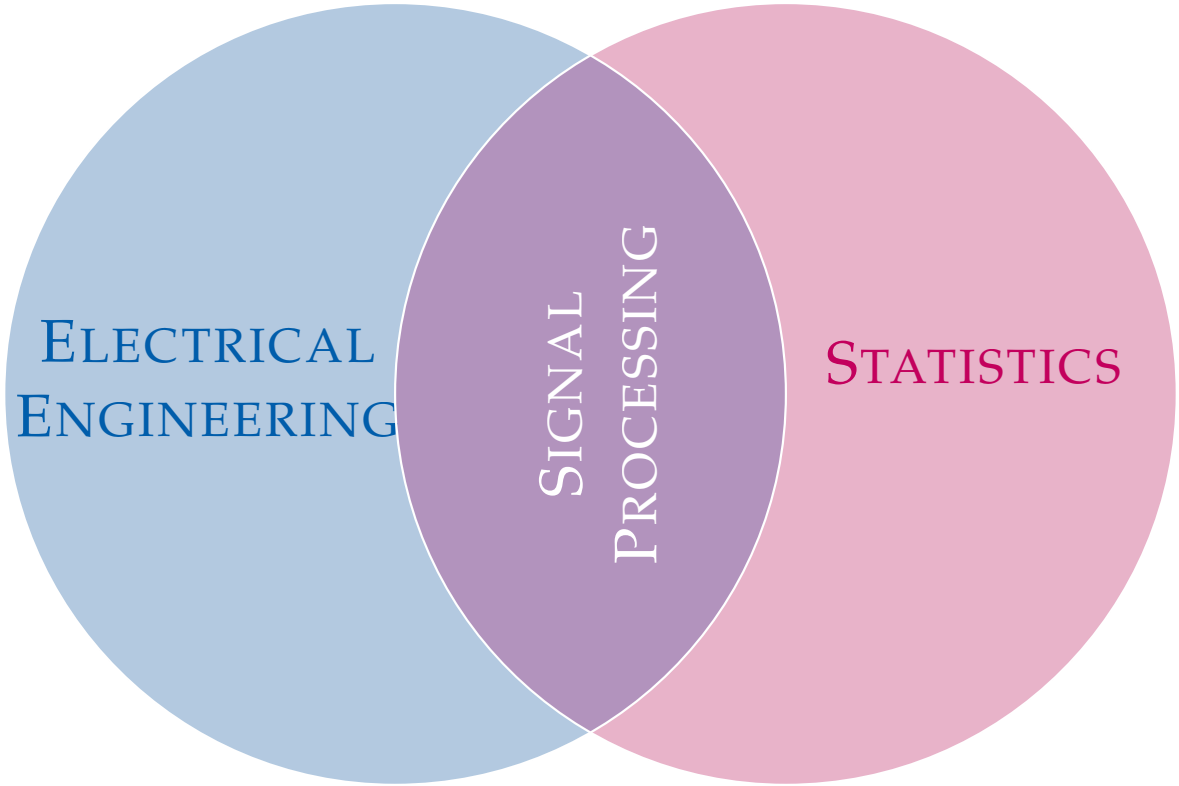
- Name + Year + Major/Intended Major + Stories

Vector open stock: `www.vectoropenstock.com`

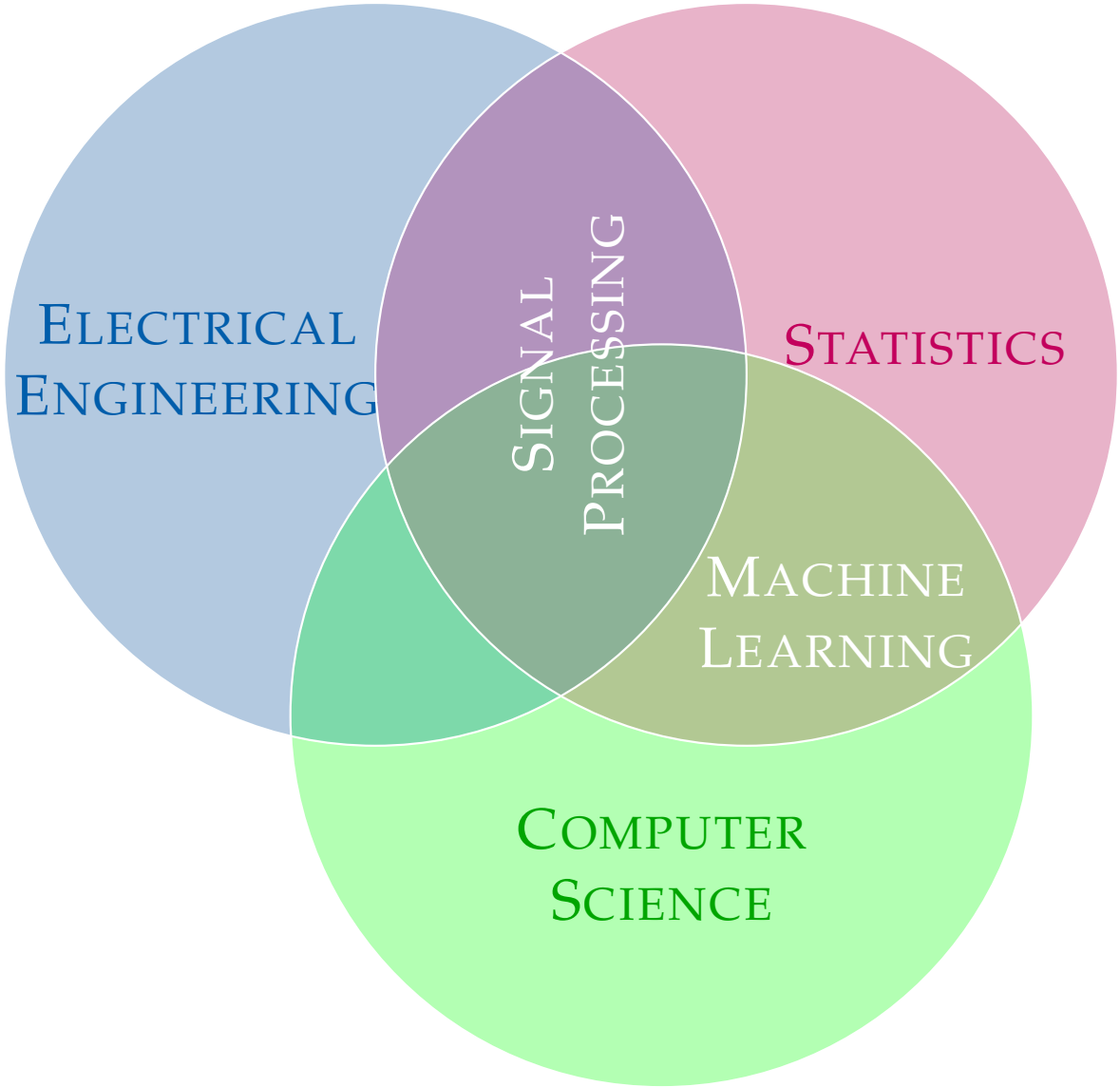
Journey Between Worlds



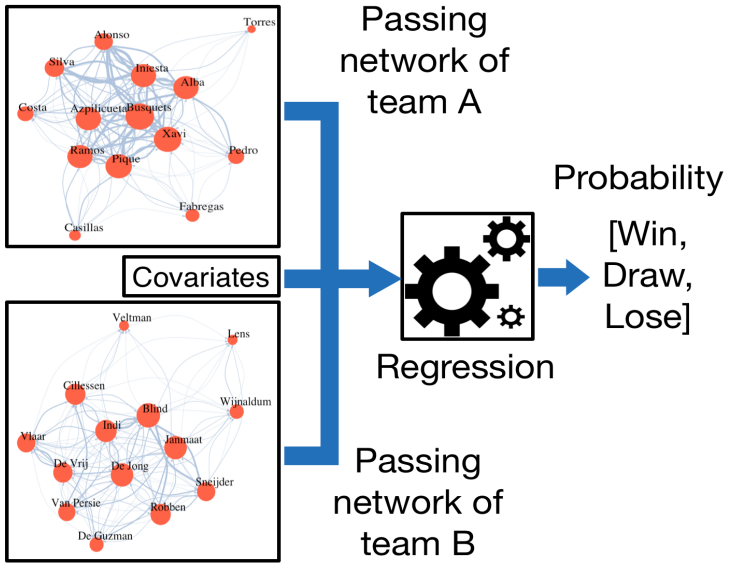
Journey Between Worlds

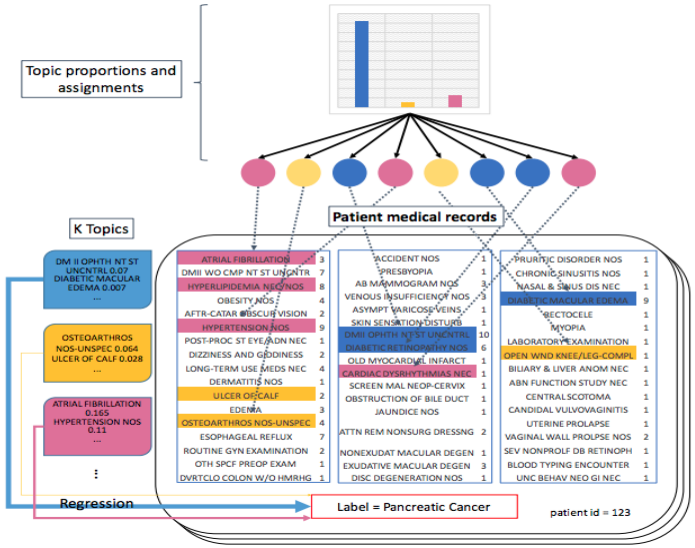


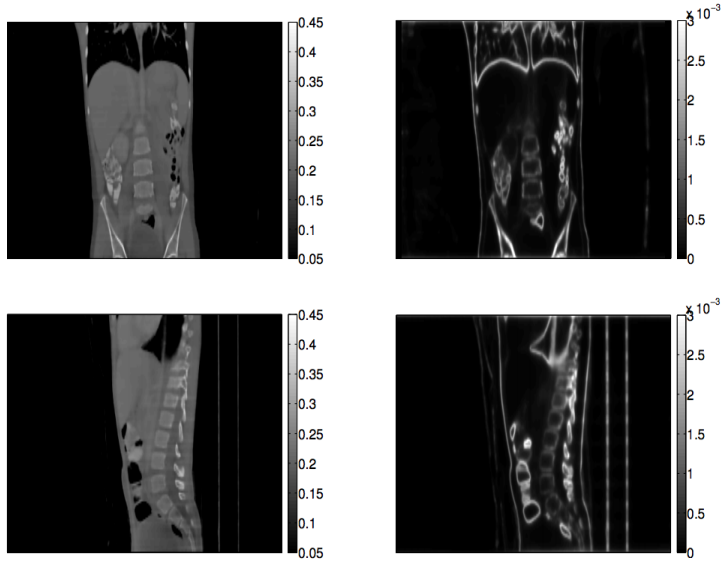
Journey Between Worlds



Work and life @Duke







Why is this course important?

The Dawning of the Age of Stochasticity

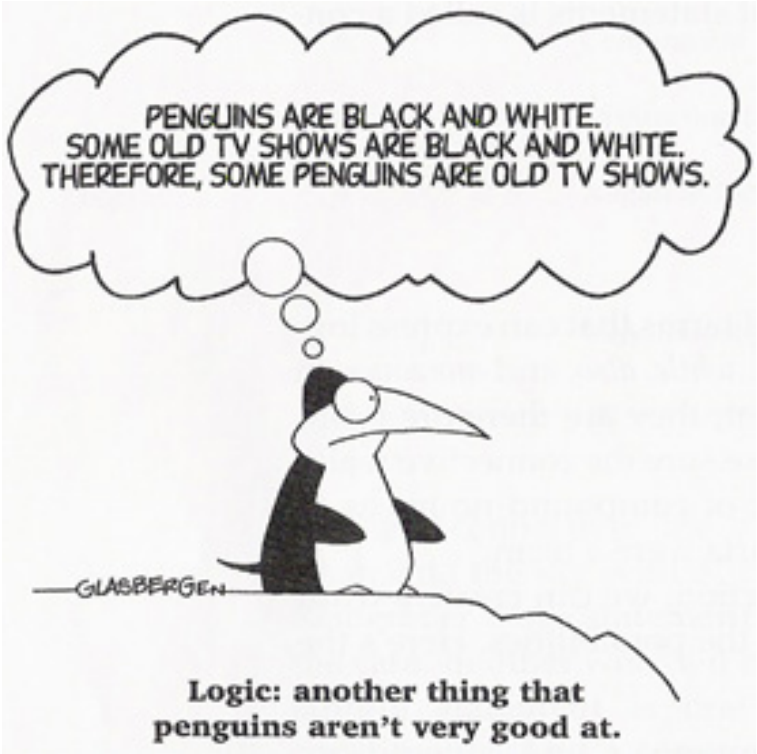
"For over two millennia, Aristotle's logic has ruled over the thinking of western intellectuals. All precise theories, all scientific modles, even models of the process of thinking itself, have in principle conformed to the straight-jacket of logic. ..... Probability theory and statistical inference now emerge as better foundations for scientific models, especially those of the process of thinking..... "

—David Mumford[1]

[1] David Mumford, *Mathematics: Frontiers and Perspectives, 2000*

Logical reasoning

- All men are mortal.
- All Greeks are men.
- Therefore, all Greeks are mortal.



https://partiallyexaminedlife.com

Probabilistic reasoning

Example: Holmes's burglar alarm[2]

1. Watson phones Holmes in his office and states the burglar alarm in Holme's house is going off. Holmes prepares to rush home.
2. Holmes recalls Watson is known to be a practical joker hence doubts his statement

   $\Rightarrow$
3. Holmes phones Mrs. Gibbon, another neighbor. She is tipsy and rants about crime, making Holmes think she has heard the alarm.
4. Holmes remembers the alarm manual said it might have been triggered by an earthquake.
5. Holmes realizes that if there had been an earthquake, it ought to be mentioned on the radio.

   $\Rightarrow$
6. Holmes turns on his radio to check.

[2] *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 1988

Troubleshooting under uncertainty

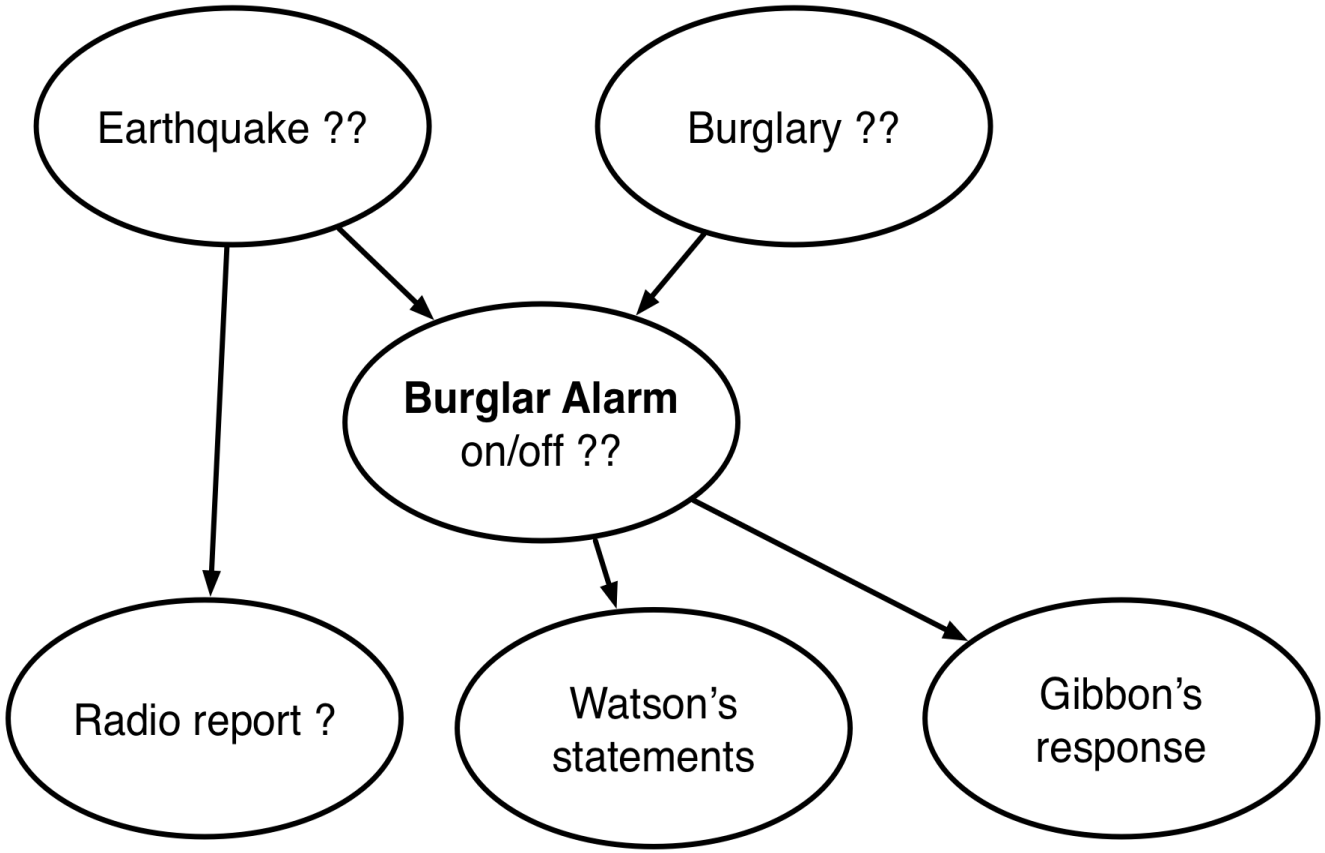


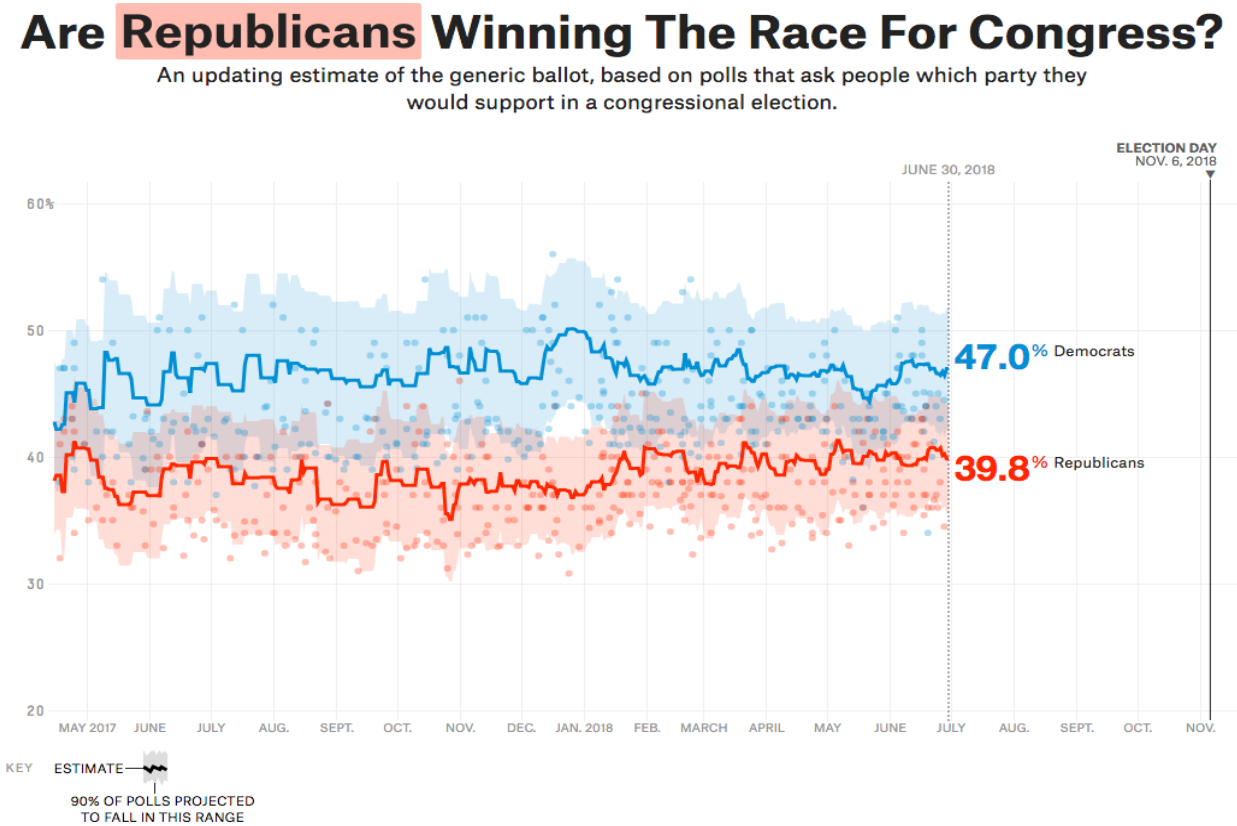
Figure from *The Dawning of the Age of Stochasticity, 2000*

- Conditional probability, Bayes' rule, statistical inference, etc.
- Troubleshooting automobile, printer, health, market, gene regulatory networks, and many others.

Why is this course useful?

FiveThirtyEight

Politics, economics, science & health, sports, culture



https://projects.fivethirtyeight.com/congress-generic-ballot-polls/?ex_cid=rrpromo

- Estimation, random variable, prediction, center and spread of distributions

S&P500 Index

Question: whether stock activity each day is independent of the stock's behavior on previous days?

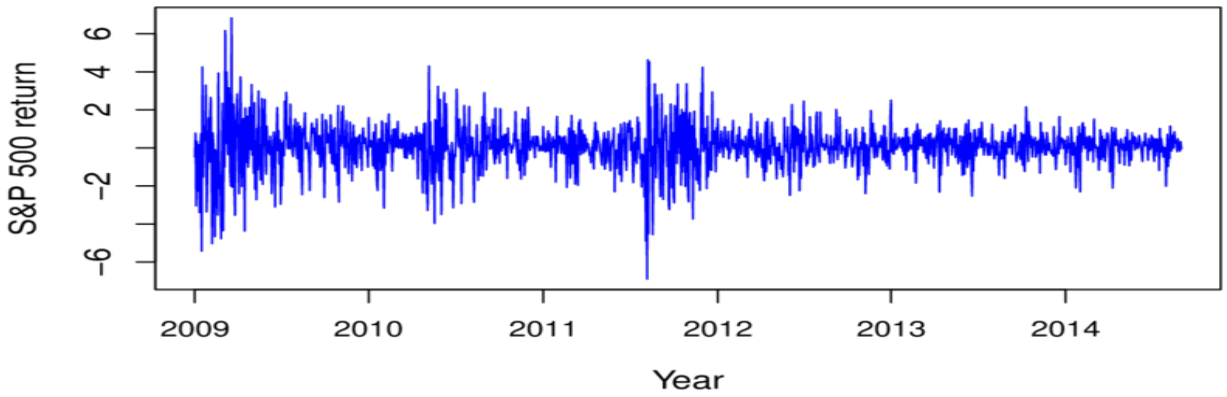


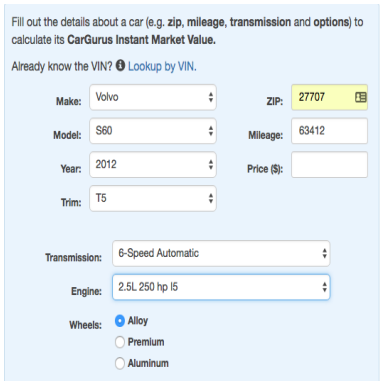
Figure from Cheng et al., *Econometric reviews*, 2018

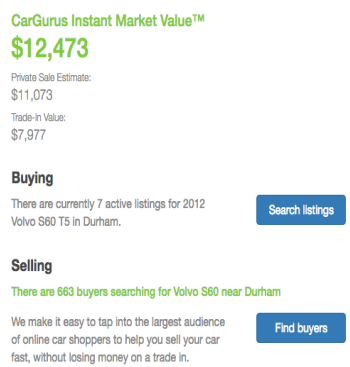
- H_0: The stock market being up or down on a given day is independent from all other days.
- H_A: The stock market being up or down on a given day is not independent from all other days.

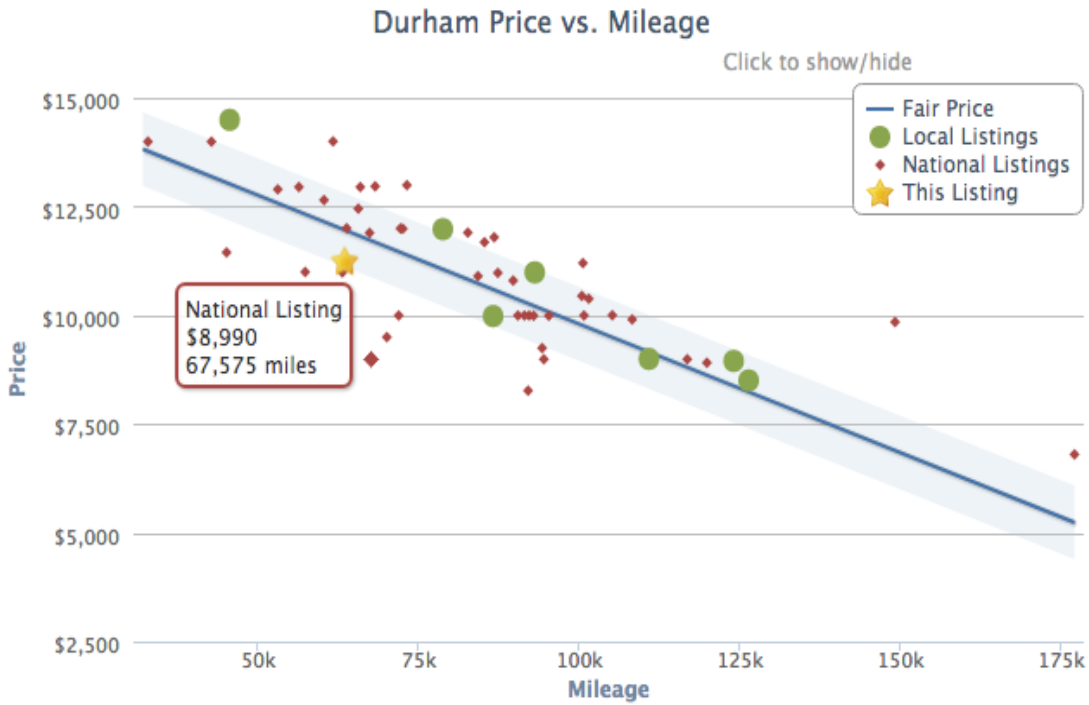
- Independence, geometric distribution, hypothesis testing, chi-square testing

Car values

Find out the Instant Market Value (or Trade-in Value) of a car







https://www.cargurus.com/Cars/instantMarketValue.action

- The method of least squares, regression, maximum likelihood estimate

Course Logistics

Synopsis

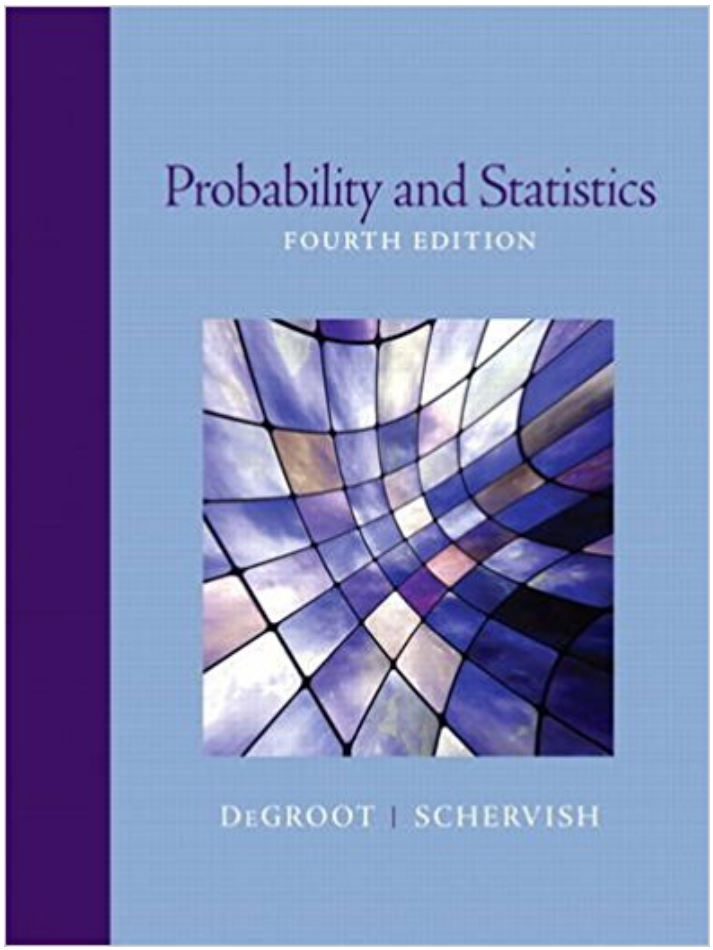
- Prerequisite: Calculus
- Probability
  1. Introduction to probability
  2. Conditional probability
  3. Random variable and distributions
  4. Expectation
  5. Special distributions
  6. Large random samples
- Statistics
  1. Estimation
  2. Sampling distribution on estimators, confidence intervals
  3. Testing hypotheses (numeric data, categorical data)
  4. Linear regression

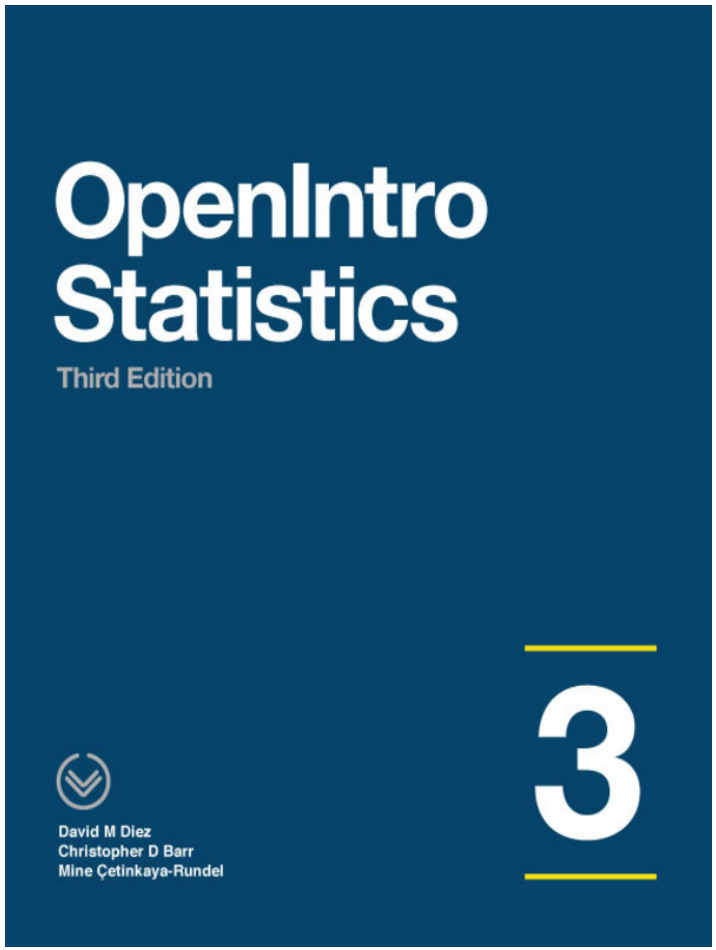
Schedule

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
7/1	7/2 Lab 1	7/3	7/4 No class	7/5 Drop/add ends	7/6	7/7
7/8	7/9 Lab 2	7/10	7/11 Lab 3	7/12	7/13	7/14
7/15	7/16 Lab 4	7/17	7/18 No Lab	7/19	7/20 **In-Class Midterm**	7/21
7/22	7/23 Lab 5	7/24	7/25 Lab 6	7/26	7/27	7/28
7/29	7/30 Lab 7 Last Day to Withdraw	7/31	8/1 Lab 8	8/2	8/3	8/4
8/5	8/6 No Lab	8/7	8/8 No Lab	8/9 Classes ends	8/10 Reading time	8/11 **Final exam**

- Lecture: Mon, Tue, Wed, Thu, and Fri 11:00 am-12:15 pm, Social Sciences 311
- Lab: Mon and Wed, 1:30 pm-2:45 pm, Social Sciences 124
- Office hours: Tues and Thurs, 4:00 pm-5:00pm, Old Chemistry 211A

Textbooks





- Textbook Reserved at Perkins (Overnight Loan)

Course Expectation

- Reading
- Lecture
- Lab
- Homework
- In class midterm
- Final exam (cumulative)

Grading Policy

- Grading:
  - Homework 30%
  - Lab report 10%
  - In-class midterm exam 25%
  - Final exam 30%
  - Class attendance 5%
- Tips for success:
  1. Be on schedule
  2. Work many problems
  3. Join a study group
  4. Calculator for homework/exam

What are the differences between probability and statistics?

Population and Samples

Considering the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. A **sample** represents a subset of the cases and is often a small faction of the population.

Anecdotal evidence

Consider the following possible response to the three research questions

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data.

1. The data only represent one or two cases
2. Are the cases representative of the population?

Probability vs. Statistics

– We will spend the first half of this course talking about probability.

– Once we have a good understanding of probability, we will switch to statistics/statistical inference.

– Essentially, probability and statistics can be thought of as inverses to each other.

– As a probabilist, you wish to make statements about data/samples/subpopulations, given what you know about the overall population.

– As a statistician, you do the opposite. Given the data you have seen, what can you say about the overall population?

Differences

[Harvey Motulsky on StackExchange]
Probability

- General $\rightarrow$ Specific
- Population $\rightarrow$ Sample
- Model $\rightarrow$ Data i.e. Given a model, what kind of data are we likely to see?

Statistics

- General $\leftarrow$ Specific
- Population $\leftarrow$ Sample
- Model $\leftarrow$ Data i.e. Given data, what kind of model is likely to have generate it?

Suppose we have a bag with a total of 100 jelly beans (some of them are red, some are green)

- The **probabilist** knows the proportion of red to green jelly beans, and want to know, for example, the probability of drawing 2 red jelly beans in a row.
- The **statistician** doesn't know the proportion of red to green jelly beans, and want to estimate it after having 2 red jelly beans in a row.

Types of questions statisticians are interested in

1. **Quantifying how precise the estimation is.** Suppose that the statistician has drawn 98 jelly beans and all of them are red. It seems clear that the estimation after drawing 98 beans will be more "precise" (in some sense) than the original estimation based on a sample size of 2 beans
2. **Deciding how many jelly beans he should drawn until he expects to achieve a sufficient precision.** Drawing jelly beans out of a bag is boring, so he might not want to draw all 100 beans and know the proportion with all certainty. The statistician might be content with estimating the proportion sufficiently well.
3. **Investigating whether the assumed probabilistic framework corresponds with reality.** Imagine that the statistician draws 30 red jelly beans in a row, but knows that the proportion of red/green jelly beans should be roughly 50%. The statistician was planning on estimating the proportion under the assumption that the jelly beans are mixed well. After seeing this, the statistician suspect that it might not be a reasonable assumption.

and many more...

Recap

- Course website: `https://shaobohan.net/courses/sta111su18/`
- Syllabus: `https://shaobohan.net/sta111/Syllabus.pdf`
- Q & A