**CS-429 Midterm Review**

1. **Review lecture slides**
2. **Review textbook: ‘’Doing Data Science”**
   1. ***Chapter 1 (what is data science)***
      1. Datafication: taking aspects of our lives and turning them into data.
      2. Explosion of data in the high-tech world.
      3. Hacking skills, math/statistics, and substantive expertise make up data science (diagram) according to Conway.
      4. Domains of data science: CS, match, statistics, machine learning, domain expertise, communication and presentation skills, data visualizations.
      5. No one person can be the perfect data scientist, so we need teams.
      6. Data scientists know how to extract meaning from and interpret data.
      7. Exploratory data analysis, combines visualization and data sense.
   2. **Chapter 2 (Statistical Inference, EDA, and the Data Science Process.)**
      1. The processes in our lives are data-generating processes.
      2. *Statistical Inference*: The overall process of going from the world to the data, and then from the data back to the world.
         1. Statistical inference is the discipline that concerns itself with the development procedures that allow us to extract meaning and information from data that has been generated by random processes.
         2. Key terms of Statistical Inference:
            1. Population – any set of objects
            2. Observation – characteristics of population (N)
            3. Sample – subset of observations (n)

Use to draw conclusions and make inference about the population.

May introduce bias. Be careful with the underlying “assumption” when you draw samples.

* + - 1. Big Data:
         1. Only when size becomes a challenge is it worth referring it as “Big”
         2. Big is when you can’t fit it all in one machine.
         3. The 4 Vs, Volume, Variety, Velocity and Value.
      2. Can N=All?
         1. It’s pretty much never all.

E.g. election night polls.

* + - * 1. It is wrong to believe either that data is objective or that “data speaks.” E.g. hiring men vs women example (ignoring causation).
      1. n = 1 (sample size of 1)
         1. In Big Data, for a single person, we can record tons of info about them.
      2. Modeling
         1. Our attempt to understand and represent the nature of reality through a particular lens.
         2. Artificial construction where all extraneous detail has been removed or abstracted.
         3. How do you build a model? What functional form should the data take.
         4. Need to make assumptions about the underlying structure of the reality
         5. Where to start? 🡪 EDA
      3. Exploratory Data Analysis
         1. This entails making plots and building intuition for your particular dataset.
         2. It is a method of systematically going through the data, plotting distributions of all variables (using box plots, plotting time series of data, … and generating summary statistics for all of them. EDA is a mindset
         3. EDA happens between you and the data. It isn’t about proving anything to anyone else yet!
         4. Namely to gain intuition about the data; comparisons between distributions; sanity checking (making sure the data is on the scale you expect, in the format you thought it should be.); to find out where data is missing or if there are outliers; and to summarize the data.
         5. First step towards building a model.
         6. Traditionally presented as a bunch of histograms and stem-and-leaf plots
         7. Exploratory = understanding of the problem.
         8. Basic tools: plots, graphs, and summary statistics.
         9. Helps with debugging the logging process. In the end, EDA help you make sure the product is performing as intended.
      5. Probability Distributions
         1. Foundation of statistical models.
         2. Normal Distribution

Bell-shaped curve

*u* is the mean and median and controls where the distribution is centered, and the parameter *ó* controls how spread out the distribution is.

* + - 1. Over fitting
         1. Term used to mean that you used a dataset to estimate the parameters of your model, but your model isn’t that good at capturing reality beyond your sample data.

1. **Review R tutorial, try out all exercises, be familiar with R basic syntax.**
2. **Review material posted on WISE.**
3. **Review commonly used statistical distributions, know basic facts about them**
4. **Review basic statistics definitions.**
5. **Understand the importance of EDA.**
6. **Understand data science as a relative new subject and data science process, data science profile.**

Questions types include answer questions, choose true or false, given R code pieces and write results.

Distributions: Poisson (over a period of time for rare events), Normal, Binomial(heads or tails), Power Law (city populations, has more of an exponential curve).