Week One: Introduction

CS 217

Probability

- Say I flip a coin. What are the odds that it will land as heads?

Probability

- Say I flip a coin. What are the odds that it will land as heads?
- The odds are ½ - we know this inherently assuming that this is a fair coin.

Probability

- Say I flip a coin. What are the odds that it will land as heads?
- The odds are ½ - we know this inherently assuming that this is a fair coin.
- Say I flip three coins. What are the odds that I will get exactly two heads?

Probability

- Say I flip a coin. What are the odds that it will land as heads?
- The odds are ½ - we know this inherently assuming that this is a fair coin.
- Say I flip three coins. What are the odds that I will get exactly two heads?
- We can use the concept of **counting** to write out all of the possibilities.

Probability

- Say I flip three coins. What are the odds that I will get exactly two heads?
- We can use the concept of **counting** to write out all of the possibilities.
- Out of **eight** equally likely outcomes, **three** of them give us exactly two heads.

HHH	**HHT**
HTH	**THH**
HTT	THT
TTH	TTT

Probability

- Another example of using counting is in rolling dice. What are the odds of getting a seven from rolling a pair of dice?

Probability

- Another example of using counting is in rolling dice. What are the odds of getting a seven from rolling a pair of dice?
- The odds are 6 out of 36 equally likely possibilities, or **1 in 6**

	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	**7**
2	3	4	5	6	**7**	8
3	4	5	6	**7**	8	9
4	5	6	**7**	8	9	10
5	6	**7**	8	9	10	11
6	**7**	8	9	10	11	12

Probability

- What are the odds of getting a seven from rolling four dice?

Probability

- What are the odds of getting a seven from rolling four die?
  - Do we really want to count this all out?

Probability

- Of course counting can get quickly cumbersome. What if we need to find the probability of getting exactly 7 heads in 10 coin flips? Or exactly 70 heads in 10 coin flips?
- We can use **probability distributions**, which we will learn throughout the course, as a means of coming up with these answers.
- These problems have defined **probability functions**. We know our coin is fair, and that our dice have a predefined set of rules. We can find the answers to our questions either through counting or evaluating their respective probability distributions.
- Counting is a good way of confirming the answers to these questions so that they intuitively make sense.
    - We can use computers to do this for us.

Probability

- In probability there is a clearly defined **experiment**.
  - We will toss exactly four die.
- There is also a clearly defined **sample space**, or range of possible outcomes.
  - If we toss four die, they can add add up to anywhere from 4 (1 * 4) to 24 (6 * 4)
- There may be an **event** that we're looking for.
  - The event that we are looking for here is that our four die add up to exactly 7.
- There is a **probability function**, or a probability of each outcome in our **sample space** occuring.
  - Each of the possible events in our sample space has a predefined probability of occuring.

Probability

- In probability there is a clearly defined **experiment**.
  - We will toss exactly four die.
- There is also a clearly defined **sample space**, or range of possible outcomes.
  - If we toss four die, they can add add up to anywhere from 4 (1 * 4) to 24 (6 * 4)
- There may be an **event** that we're looking for.
  - The event that we are looking for here is that our four die add up to exactly 7.
- There is a **probability function**, or a probability of each outcome in our **sample space** occuring.
  - Each of the possible events in our sample space has a predefined probability of occuring.
- Unfortunately, the real world is usually not that simple.

Statistics

- A **statistic** is anything that can be computed from collected data.
- **Probability** deals with the likelihood of predicting future events, while **statistics** involves the analysis of the frequency of past events
- Statistics involves the collection of data, its subsequent description, and its analysis, which leads to the drawing of conclusions.
- **Probability** is to **Statistics** as **Calculus** is to **Physics**
- To be an effective statistician (or data scientist), you must have an understanding of the essential probability concepts, models and distributions

Descriptive Statistics

- **Descriptive Statistics** involve **describing the data** without drawing conclusions.
- Aaron Judge got up to bat in 2018 498 times. He had 115 hits, 22 doubles, 27 home runs, and 152 strikeouts.
- These are all **descriptive statistics** because they are data points that don't infer anything about his performance.

Summary Statistics

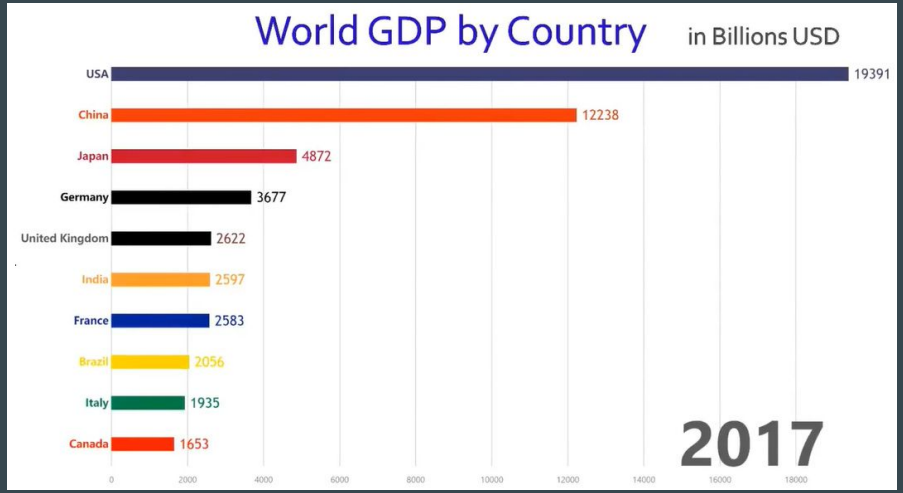
- **Summary Statistics** summarize a set of observations, in order to communicate the largest amount of information as simply as possible.
- A **summary statistic** about Aaron Judge may be that his batting average last year was .278
- This is a concise way of understanding how good he was, with the caveat that this doesn't tell us everything about his performance.

Summary Statistics

- What are other examples of summary statistics (in baseball, economics, school, life...)?

Summary Statistics

- Some examples may be:
  - Your GPA as a measure of your academic performance.
  - Approval rating as a measure of presidential performance.
  - GDP as a measure of economic production from a country.
  - Blood pressure as a measure of one's health.
  - Gini Index as a measure of inequality within a country.



What are Statistics?

- **Inferential Statistics** involve **making inferences** on a dataset and **drawing conclusions** from it about the larger world.
- We can make a hypothesis and then use data to provide evidence for or against this hypothesis.
- We need to ensure that we are clear in the definitions of our hypotheses.
- We also need to ensure that the data is representative of the larger world.

Inferential Statistics

- A poorly designed experiment will produce poor quality data.
- A famous example of a **biased sample** is that in 1936, a popular magazine sent out a poll to 10 million of their readers asking them who they were going to vote for in the upcoming election - Republican Alf Landon or Democrat Franklin Roosevelt
- They received two million ballots showing that Landon would get 57% of the votes in the election.
- Of course there was no President Alf Landon - most subscribers of the magazine were affluent and not representative of the general population. The sample was biased and thus useless even given the huge sample size.

Applications for Statistics

- Weather Predictions
- Economic reporting
- Political polling
- Sports
- Box Office Reporting
- Marketing
- Gambling
- Genetic Testing
- Insurance
- TV Ratings
- Finance
- Social Media Algorithms
- Streaming Algorithms
- Medical Research

Applications for Statistics

- How can we catch schools that are cheating on their standardized test?
- How does Netflix know what kind of movies you like?
- How can we figure out what substances of behaviors cause cancer in humans given that we cannot conduct cancer-causing experiments on humans?
- Does praying for surgical patients improve their outcomes?
- Is there really an economics benefit to getting a degree from a highly selective college or university?
- What is causing the rising incidence of autism?

Risk Assessment

- How does a casino know that it can successfully pay out game winners (and keep the games competitive enough that people will want to play), but remain profitable?
- How does Geico know how much to charge you for auto insurance?
- How can you build a stock portfolio that will profitable without being too risky?

Expected Value is a big tenent of risk assessment.

Causal Relationships

- **Does smoking cause cancer?**
- Scientifically you'd want to create a controlled experiment of smokers and non-smokers and measure the rate of cancer in each group after a period of time.
- You can see how this would be expensive and totally unethical to pull off.
- You can simply measure the rate of cancer in existing smokers vs. non-smokers, but keep in mind that there are other variables which may not be clear at first.
- For instance people who tend to smoke may have other lifestyle habits that affect their well-being.
- Identifying causal relationships is extremely hard - we will point out some common mistakes with this later in the course.

Lies, Damned Lies, and Statistics

- Who is a better baseball player?
  - *What metric are you using to compare?*
- How has the health of America's middle class changed in the past twenty years?
  - *How do you define 'middle class' and 'health'?*
- IQ is correlated with income, thus people make more money because they're smart.
  - Is correlation equal to causation here?
- Even with good intentions, statistics requires clear definitions of imperfect data.
- Nevermind that people will use statistics in bad faith **ALL THE TIME.**

Welcome to CS 217!

- What is the goal of this course?
  - To introduce you to the core concepts of probability and statistics
- How will you learn in this course?
  - Via hands-on-learning - the course takes a computational and applied approach to our topics
- What language will we be using?
  - The class will be administered entirely in Python. If you've never used Python before, don't worry! No prior knowledge is required.
- How will we spend our time during class?
  - Class will be split between lectures and hands-on group work, with occasional quizzes, announced and unannounced, to check for understanding.

Course Agenda

- Descriptive Statistics
- Discrete and Continuous Distributions
- Normal Distribution and Central Limit Theorem
- Hypothesis Testing
- Relationships Between Variables & Regression

Course Objectives

By the end of the course, students should be proficient at:

1. **Single Variable Explorations**: Examine a single variable, understand its underlying distribution, and choose the appropriate summary statistics for it.
2. **Pair-Wise Exploration**: Identify possible relationships between variables and compute correlations and linear fits.
3. **Estimation and Hypothesis Testing**: Understand the following three questions when reporting statistical results: 1) How big is the effect? 2) How much variability should we expect if we run the same measurement again? 3) Is it possible that the apparent effect is due to chance?
4. **Visualization**: Use data visualization as a tool for examining data and communicating results

Grading

	Weight
Final Project	25%
Midterm Exam	25%
Final Exam	25%
Homework/Quizzes	15%
Participation	10%

Tools

- **Python** for Data Analysis
  - Almost everything we do in the class will only use four or five packages
- **Binder** for executing Python in the cloud
  - We will use this as a resource to complete in-class assignments and homework.
- **Github** to host all class material
  - Available at https://github.com/CSC217/fall_2019
- **Slack** for class communication
  - Slack will be the main channel for administrative updates, but you are also encouraged to use it to communicate with each other for collaboration.
- **Kahoot** for informal, in-class quizzes
  - Kahoot is an app that lets you create and distribute quizzes for a group setting

Textbooks

- *Think Stats: Exploratory Data Analysis in Python*, Allen B. Downey, Second Edition
  - A computational and Python-focused introduction to key statistical concepts
- *Think Bayes: Bayesian Statistics Made Simple*, Allen B. Downey, First Edition
  - We will touch on a few topics in this book - overall a good follow-up to this class
- *Introduction to Probability and Statistics*, Jeremy Orloff and Jonathan Bloom
  - Supplementary, in-depth notes related to the class materials.
- Readings will be assigned from these texts each week, along with readings from across the web.

About Me

- I'm about to start as a Data Scientist at Stash, a financial technology helping people learn how to invest.
- Previously, I worked as a Data Scientist at 360i, an advertising agency, since late 2017.
- I have a BA in Economics from Boston University and an MS in Applied Statistics from Penn State University.

About Me

- I'm working with the NYC Tech-In-Residence Corps to teach you about concepts and tools we use in the workplace.
- This is why we're using Python and focusing on the applied end of statistics - I want you to see how it's useful from a professional perspective rather than looking up Z-tables in a textbook and talking about counting colored balls from an urn (though we may do a bit of that)