2019

Lecture: Probability and Statistics - Introduction - Week One

Evan Agovino
CUNY City College

NYC Tech-in-Residence Corps

Week One: Introduction

•••

CS 217

What are Statistics?

- Simply put, statistics is the measuring and interpretation of data.
- It involves the collection of data, its subsequent description, and its analysis, which leads to the drawing of conclusions.
- Statistics are used in numerous aspects of our lives whether we realize it or not.
- What are some examples of the use of statistics in day-to-day life?

Applications for Statistics

- Weather Predictions
- Economic reporting
- Political polling
- Sports
- Box Office Reporting
- Marketing
- Gambling

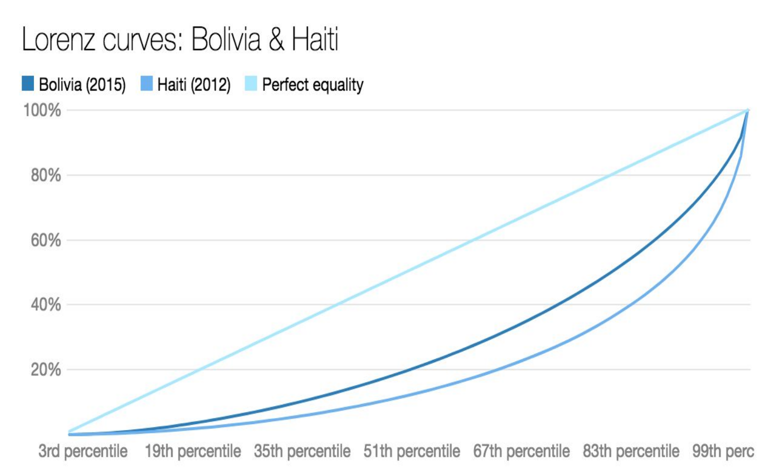
- Genetic Testing
- Insurance
- TV Ratings
- Finance
- Social Media Algorithms
- Streaming Algorithms
- Medical Research

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?

Description and Comparison

- Summary statistics exist that take an incredibly complicated topic and boil it down to one number
- **Batting Average** as a measure of how good a baseball player is
- **GPA** as a measure of how good a student you are
- **Gini Index** as a measure of income inequality across a country



Each of these metrics has a tradeoff, they are helpful as a concise way of sharing incredibly complex information, but can only tell so much of the story

Inference

- How many bodegas are in New York City?
- Who will win the Democratic primary?
- How much money did The Avengers: Infinity War make in the box office on its first weekend?

Often is it expensive or impossible to calculate these fully, but we can get a representative sample and extrapolate an estimate of the full data.

We need to be careful though and make sure our representative sample is **large enough** and **unbiased**.

Inference

- 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.

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?

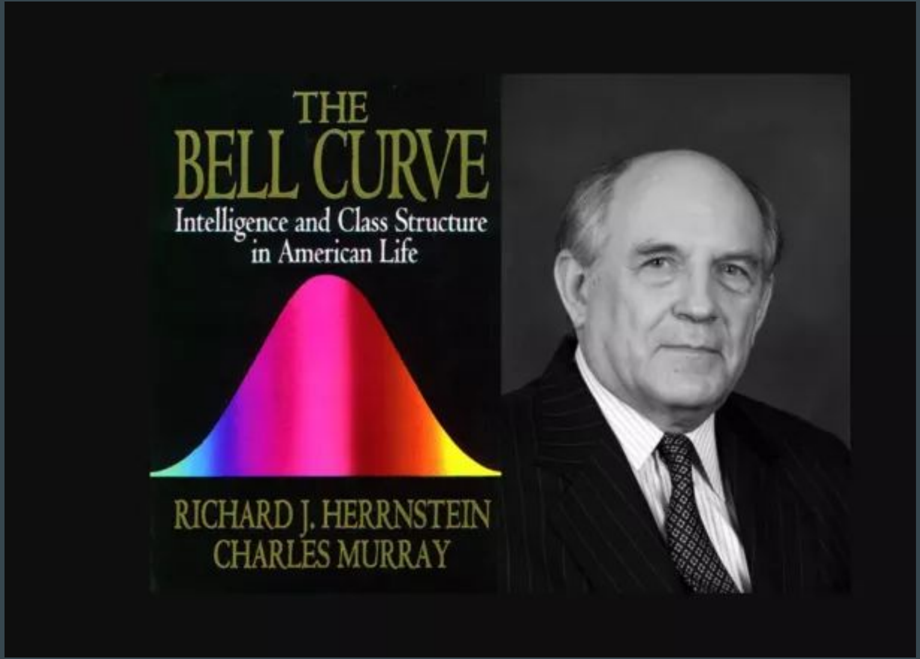
Expected Value is a big tenent of risk assessment.

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 in this analysis later in the course.

Lies

- 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'?*
- Even with good intentions, statistics requires clear definitions of imperfect data.
- Nevermind that people will use statistics in bad faith **ALL THE TIME**.



Why Stats?

- Summarize huge quantities of data
- Make better decisions
- Answer important social questions
- Recognize patterns in behavior
- Evaluate the effectiveness of politics, programs, and other procedures
- Catch people using statistics nefariously

What are Statistics?

- There are two different types of data analysis: **descriptive statistics** and **inferential statistics**
- **Descriptive Statistics** involve **describing the data** without drawing conclusions
- Say you are studying the effect of a new medicine that is set to lower a patient’s fever. You have a dataset of twenty patients.
- With descriptive statistics, you can find out the average temperature at which the fever was reduced, along with the total range of temperatures in which the fever was reduced and the variability in which temperatures for the fever were reduced.
- You are not trying to prove or disprove a specific hypothesis, but you have learned more about your data by exploring its tendencies.

What are Statistics?

- There are two different types of data analysis: **descriptive statistics** and **inferential statistics**
- **Inferential Statistics** involve **making inferences** and **drawing conclusions** about the data in a dataset.
- Say you want to know which types of car maintenance lead to the biggest increase in fuel efficiency.
- You take several observations of different inputs - tire pressure, oil changes, quality of gasoline, and outside temperature - and then use statistical tools to determine which of these inputs has the most effect on the car’s fuel consumption - and which didn’t.

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
- Basic Probability
- Random Variables and Distributions
- Normal Distribution and Central Limit Theorem
- Estimation and Confidence Intervals
- Hypothesis Testing
- 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
Group 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/spring_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

- *Introduction to Probability and Statistics for Engineers and Scientists*, Sheldon M. Ross, Third Edition. Available for free online.
    - This is the mathier book but a very good comprehensive reference for the class.
- *Think Stats: Exploratory Data Analysis in Python*, Allen B. Downey, Second Edition. Available for free online.
    - This book has a more layman's approach, with examples and code intertwined.
- Readings will be assigned from each of these books each week, along with readings from across the web.
- How you ingest the readings is up to you, of course. I'd recommend reading the material from Think Stats first to get a simple overview of the material before diving into Introduction to Probability and Statistics.

About Me

- I’m currently a Data Scientist at 360i, an advertising agency. I’ve been there since late 2017.
- Specifically I work in the programmatic department, helping our clients optimize their bids on targeted display, video, and audio ads.
- 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)

NYC Tech-In-Residence

TIR CORPS

EST. 2017