ENGRD 2700:
Basic Engineering Probability & Statistics

Johannes Wissel

School of Operations Research & Information Engineering
Cornell University

Spring 2024

Building on notes written by Kenneth Chong, Shane Henderson, Jefferson Huang, Sid Resnick, David Matteson, Dawn Woodard, Yudong Chen

What's "Probability & Statistics"?

It's the study of *models* and *data* to help you:

- *understand* a system better, and/or
- make a *decision*.

It plays an essential role in the *scientific method*.

W. Edwards Deming (Engineer & Statistician):

> *In God we trust; all others must bring data.*

Who Else Cares?

Laszlo Bock (Senior VP of People Operations, Google):

> *I took statistics at business school, and it was a transformative experience. Analytical training gives you a skill set that differentiates you from most people in the labor market. (NYT, 2014)*

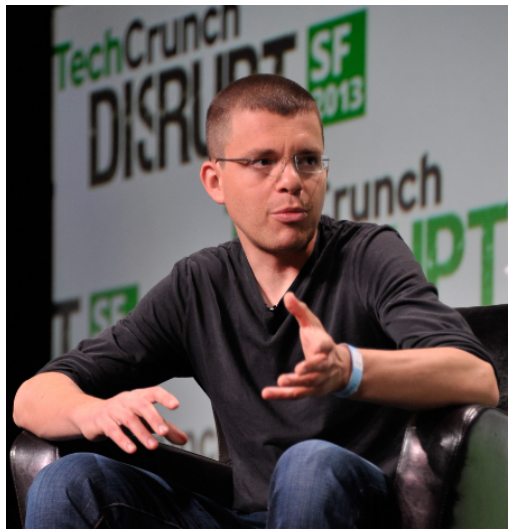
Hal Varian (Chief Economist, Google):

> *I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding. (NYT, 2009)*

Who Else Cares?

Max Levchin (Paypal Cofounder):

> *I am not much given to regret, so I puzzled over this one a while. Should have taken much more statistics in college, I think. (ASA, 2010)*



Why Care?

Probability & statistics is a prerequisite for **data mining** and **machine learning**. It helps you *reason under uncertainty.*

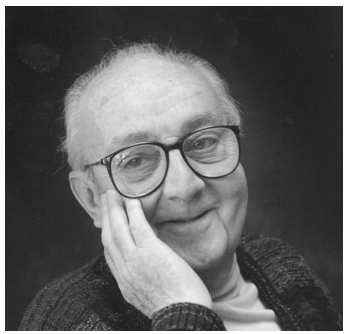
- **Civil Engineering:** Is the dam high enough?
- **Computer Science:** Is your new email message spam?
- **Energy:** Is the wind farm cost-effective?
- **Finance:** Is the bank's reserve large enough?
- **Mechanical Engineering:** Will the part fail during the next five years?
- **Medicine:** Is the drug safe and effective?
- **Operations Research:** Where should ambulances be based?

What's this course about?

1. **Probability:** Building models of **random phenomena**.
   - A phenomenon is *random* if its outcome can't be predicted in advance, e.g.:
     - Which side a coin lands on;
     - Tomorrow's stock price;
     - The result of an upcoming election.

George Box (Statistician):

All models are wrong; some models are useful.

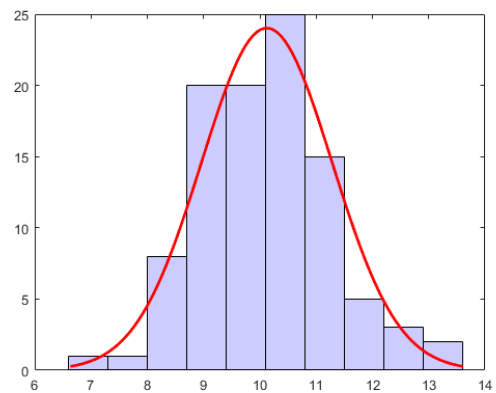


What's this course about?

2. Statistics: (1) *Organizing & summarizing* data, and (2) using the data to *draw conclusions*, e.g.:

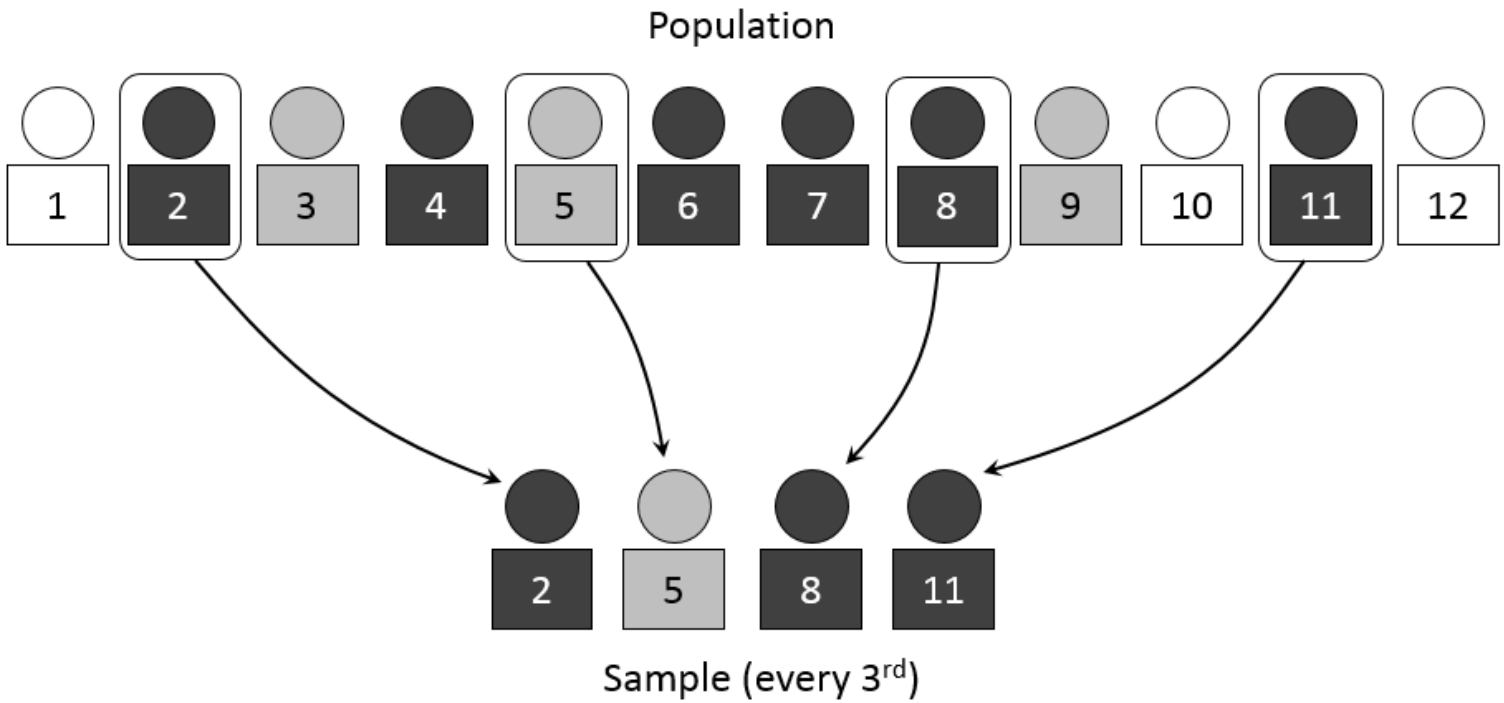
- Based on the outcomes of some tosses, is the coin fair?
- Based on the stock's price in the past, will it go up tomorrow?
- Based on survey data, who will win the election?

Statisticians use probability, e.g. fit probability models to data (a *sample*) from a *population*, and use the models to draw conclusions.



Population vs. Sample

- **Population:** a well-defined set of items you're interested in.
- **Sample:** a well-defined subset of a population (often used to study that population).
- **Observation:** an individual measurement from a sample.



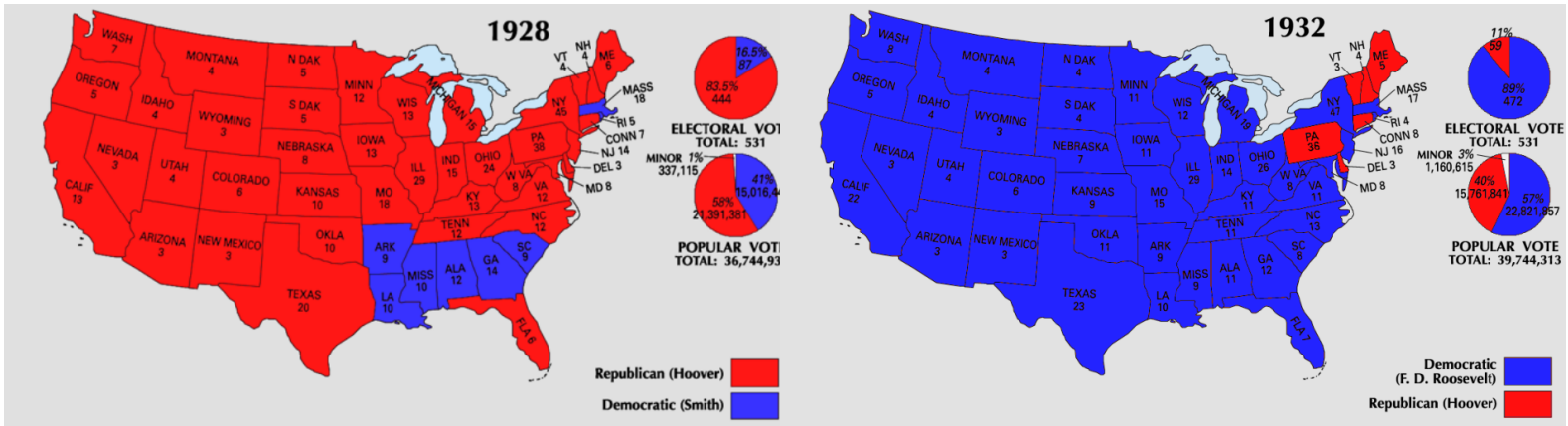
Population vs. Sample: Examples

- All students taking ENGRD 2700 (population), vs. all sophomores taking ENGRD 2700 (sample);
- All U.S. universities (population), vs. all Ivy League universities (sample);
- All voters in the U.S. (population), vs. all college-educated voters (sample);

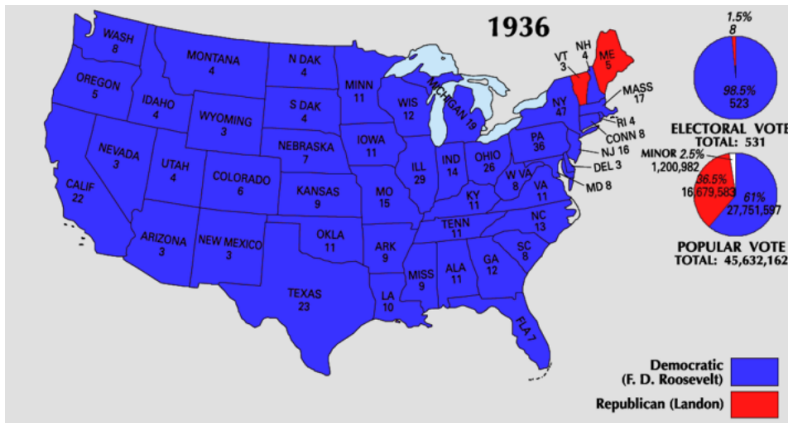
When using a sample to learn about a population, make sure the sample is *indicative of the population*.

Example: When sampling goes wrong

The Literary Digest correctly predicted the winner of all the presidential elections between 1916 and 1932.



They blew it big time in 1936. **Why?**



Example: When sampling goes wrong

Answer: They used automobile registration lists and telephone directories to sample voters.

- Before 1936, the upper & working classes had similar political opinions.
- In 1936, the upper class was mostly Republican, and the working class was mostly Democrat.
- What kinds of people tended to own automobiles/telephones in 1936?
- Luckily, today we're more sophisticated and never make mistakes...

The 2016 Presidential Election

- FiveThirtyEight.com gave Trump a 28.6% chance on the eve of the election
- Most (not all) other sites predicted a much lower chance
- What did the analyses miss?
- Did those polled represent those who voted?
- Did those polled tell the truth? (Bradley effect)
- This was a close one, e.g., Trump won Wisconsin by 27,000, and 300,000 registered voters lacked strict voter ID.

Statistical Analysis: Four Stages

1. **Formulate** clear, answerable questions about your precisely defined population.
2. **Collect** data that will help answer your questions, using a well-chosen sampling scheme and experimental design.
3. **Explore** the data, using
   - graphics;
   - descriptive statistics (e.g. mean, median, variance, quantiles).
4. **Analyze** the data to *draw conclusions* about the population.

Types of Samples

1. **Simple random sample of size n:** Each subset of size n of the population is equally likely to be chosen.
2. **Stratified random sample:** Collect random samples from sub-populations of the population.
   - e.g., sample 700 Cornell undergrads by sampling 100 undergrads from each college.

The data from a sample can be *real-valued* (e.g., pollution concentration), *integer-valued* (e.g., defect counts), or *categorical* (e.g., makes of cars).

Exploratory Data Analysis vs. Formal Inference

- **Exploratory data analysis (EDA):** Summarize main characteristics of the data (e.g., via plots, summary statistics such as the mean, median, variance). This can be used to:
  - get a better (qualitative) understanding of the data;
  - formulate questions about the data.

- **Formal inference:** Draw scientific inferences about the population from the data (e.g., formally test hypotheses).
  - Clinical trials (most famous: for Salk et al.'s polio vaccine in the 1950s).
  - The hypotheses might come from EDA.

Big data sets are common!

- Retail transactions (e.g., Walmart, Amazon);
- high-frequency (e.g., fractions of a second) financial data;
- ambulance call records;
- astronomical data (e.g., images from telescopes)
- dating profiles (e.g., Match.com, OKCupid)
- Uber ride information
- tweets on Twitter
- photos (e.g., Google, Facebook, Yelp)

Data scientists try to extract useful information from large data sets; see e.g., https://www.kaggle.com/.

Summary

- This course is about *probability & statistics*.
- Probability & statistics are important for *making decisions under uncertainty*.
- Most statistical analyses follow *four stages*:
  - Define population and question.
  - Choose sample.
  - Exploratory data analysis (play & plot).
  - Formal inference (prove).

Syllabus

On the course canvas page, go to "Modules"

https://canvas.cornell.edu/courses/60732/modules