Big Data and Security

Jeffrey Borowitz, PhD
Lecturer
Sam Nunn School of International Affairs

Random Variables, Samples, and the Laws of Probability

Sampling

- The way we think about data to analyze is as a **sample**
  - This is the only game in town
- We want to look at the distribution of a bunch of random variables
- So we observe the outcomes of that random sample a lot of times
  - If you want to understand dice behavior, roll them a bunch to get dice outcomes
  - If you want to understand how people's backgrounds determine their wages, survey a bunch of people
  - If you want to analyze behavior of a customer, survey a bunch of customers
- We usually assume that observations are **independent**
  - This isn't strictly necessary, as long as you know specifically how things are independent.
  - For example, the Current Population Survey, which measures unemployment in the US.

Creating Samples: Draws from a Population

- **Population** is the group of units from which each observation is drawn
- This term naturally applies to thinking about people or groups of animals
  - The population is all the people in e.g. the United States
- When we roll dice, what is our population?
  - It’s the abstract random variable which generates die rolls
  - When you think about it, this is what we always want to know about
    - The abstract die-roll like even which generates flu?
- Our **population** in a lot of cases makes more sense to think of as a **data generating process**
  - We try to use the **sample** to decide about the data generating process
  - If we know the data generating process, we can make predictions

Georgia Tech

Modeling and Sampling

- So we typically want to understand something about the random variable X (or die rolls) from seeing a bunch of observations
- What does a model mean here?
  - Typically, we think E[X] (the expectations of a random variable) depends on some factors
  - E.g. the expectation of how much money I make depends on how long I went to school
  - The expectation of how many flu cases depends on how many times people Google flu related symptoms

Sampling and Big Data

In this framework,
what does it meant to have “all” the data?

Sampling and Big Data

- In this framework, what does it meant to have “all” the data?
  - Nothing! What would it mean to have all dice rolls?
- A possible exception: what if you only care about the model for people you have data for.
  - You have a set of customers
  - Your model has Y as whether a person buys a particular good, and X as whether you show them a blue or red website
  - If you never want to show your website to anyone else, you know the effect of website color on buying for the population

Georgia Tech

Law of Large Numbers

- If you have a big sample of many draws from X, their average will eventually get arbitrarily close to E[X]
- This is true for the expectation of X in the population being sampled
  - If you survey e.g. only landline phones, how will your survey cover people without phones?

What Does the Law Of Large Numbers Mean for Us?

- Pretty much all statistics are based on the law of large numbers
  - The way you do this is use math to rewrite the model as a sum of random variables
- Law of large numbers says the average will converge to the expectation
- So what does the expectation mean?
  - We're going to converge to the population average
  - For dice, the population is really the process that's the population
  - In general, an important question to ask is what is the population you're learning about with statistical analysis.

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

$$0 = \sum_i \varepsilon_i = \sum_i y_i - \alpha - \beta x_i$$

Central Limit Theorem

- If you have a a bunch of big samples of many draws from X
- Then the average of each of these big samples is another random variable (functions of random variables are random variables)
- And the distribution of these sample averages is **normally distributed** with:
  - Average: E[X]
  - Variance: var(X)/N where Ns the number of observations in each sample
  - So variance decreases with N

Simplified Implications of LLN and CLT

- LLN says that if our models are right, and we use enough data, our predictions will be right on average
- CLT says that if our models are right, we also know how accurately we know our predictions

Lesson Summary

- Random samples are utilized to represent the population and allow for a better understanding of a variable X
- The Law of Large Numbers states that the larger the sample and the more “draws” done, the closer the average will be to the expectation, E(x)
- Central Limit Theorem puts the ”draws” into a normal distribution with E(x) as the average and var(X)/N as the variance (N is number of samples)