Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Probability Part 1:
Random Variables

Probability

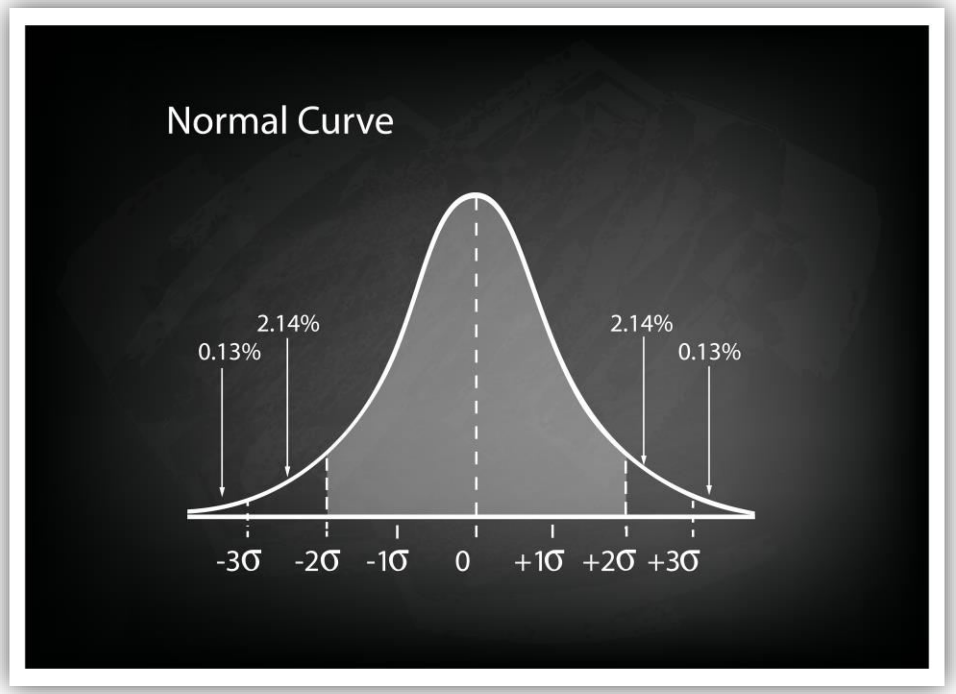
- A random variable
  - Has a **set of potential outcomes**
  - Has a **probability distribution** over outcomes
- Example: the outcome of a dice roll as a random variable
  - Set of potential outcomes: {1, 2, 3, 4, 5, 6}
  - Probability distribution f (DICE VALUE)
  - Probability f (DICE VALUE = x) is the chance that the die roll is x

$$\text{f (DICE VALUE)} = \begin{cases} \frac{1}{6}, \text{DICE VALUE} = 1 \\ \frac{1}{6}, \text{DICE VALUE} = 2 \\ \frac{1}{6}, \text{DICE VALUE} = 3 \\ \frac{1}{6}, \text{DICE VALUE} = 4 \\ \frac{1}{6}, \text{DICE VALUE} = 5 \\ \frac{1}{6}, \text{DICE VALUE} = 6 \\ 0, otherwise \end{cases}$$

Georgia Tech

Probability Distribution: Graphically

Probability Distributions: Could be Continuous Too



Why Do We Care?

- The goal is **not** to introduce needless math here
- But probability is the basic framework which is used to model data that we do care about
- Flu trends:
  - Random variable: number of flu cases in a state in a week
  - Potential outcomes: 0, 146, 10, 000, 000, etc.
  - Probability distribution: ???
- Consumer modeling
  - Random variable: did I purchase a particular book from Amazon?
  - Potential outcomes: yes, no
  - Probability distribution: ???
- Threat modeling
  - Is a particular individual a threat?
  - Potential outcomes: yes, no
  - Probability distribution: ???

Georgia Tech

Random Variables Have a Mean

- What is the mean (or average) value of the roll of a die?
  - Think about a simpler case: a coin flip where Heads counts as 1, Tails counts as 0:

$$\frac{1}{2}\cdot(1)+\frac{1}{2}\cdot(0)=\frac{1}{2}$$

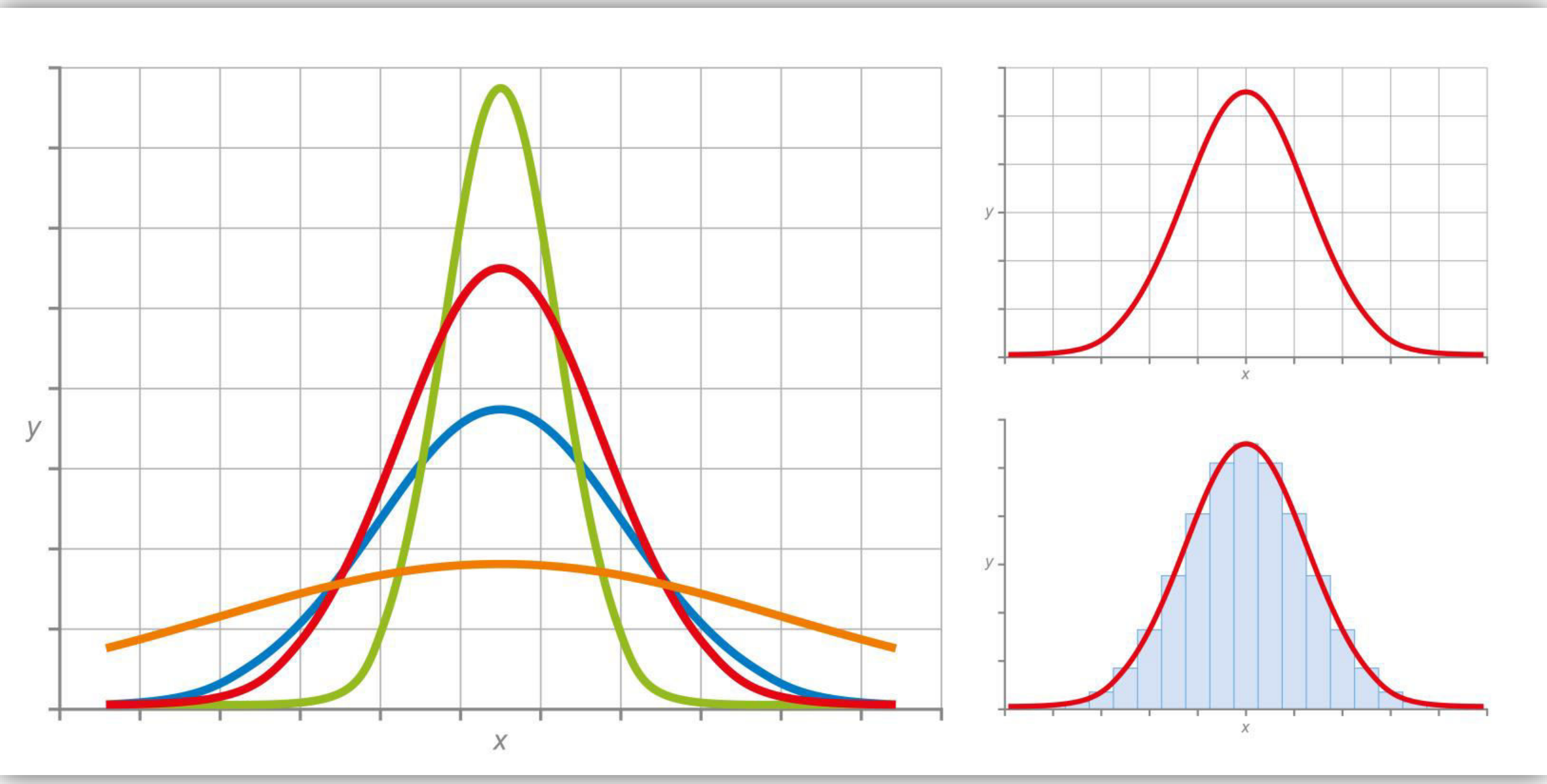
  - The analogous calculation for the die roll:

$$\frac{1}{6}\cdot(1)+\frac{1}{6}\cdot(2)+\frac{1}{6}\cdot(3)+\frac{1}{6}\cdot(4)+\frac{1}{6}\cdot(5)+\frac{1}{6}\cdot(6)=\frac{7}{2}=3.5$$

- The mean is where the graph of the probability distribution would balance

Random Variables Have a Variance

This is a measure of how spread out the distribution is:



Random Variables Have Quantiles

- The p-th quantile (called q) of the random variable X is the value of x where there is a p percent chance of X having a value at or below q

$$prob(X < q) = p$$

- Example: the 83rd quantile of rolling a die is 5
  - 5/6 of the time (83%), we role a number 1-5
- Example: the 60th percentile of household income in the US is about $80,000
  - 60% of households make less than this.
  - Or: if we draw a random household from the US, there is a 60% chance that the household makes less than $80,000

Expectation Operator

- In addition to calculating the mean and variance, we can calculate the expected value of any function of the random variable X
  - Called E[g(X)] for the function g

$$E[f(x)] = \frac{1}{6}f(1) + \frac{1}{6}f(2) + \frac{1}{6}f(3) + \frac{1}{6}f(4) + \frac{1}{6}f(5) + \frac{1}{6}f(6)$$

  - The intuition is: what is the average value of the function, when g depends on the outcome of the dice roll
  - Example: a car salesperson wants to calculate her bonus
    - If she sells no cars, she gets $1,000 (prob. 1/4)
    - If she sells 1-5 cars, she gets $4,000 (prob. 1/2)
    - If she sells over 5 cars, she gets $10,000 (prob.1/4)
    - The expectation of her bonus is:

$$1000\left(\frac{1}{4}\right) + 4000\left(\frac{1}{2}\right) + 10000\left(\frac{1}{4}\right) = 4750$$

Lesson Summary

- Probability is the likelihood of a certain outcome
- Random variables:
  - Have a mean, which is located where the distribution curve would balance
  - Have a variance which is represented in how spread out the distribution is
  - Functions of random variables are random variable
- We'll use random variables to model data