DATA 252/DATA 551
Modeling and Simulation

Lecture 1: Discrete Random Variables

January 13, 2020

Random variables and sample spaces

Consider a random experiment like tossing a fair coin. This is a random experiment because the outcome is unknown. We can represent this unknown outcome with X, which is called a random variable.

Sample space: the sample space is the set (i.e., collection) of *all* possible outcomes.

- Sample space of tossing a fair coin: $\{H, T\}$

Random variables and sample spaces

{TTT, THT, ..., } ⇒ sample space Ω

- Suppose you toss a fair coin three times. Let X be the number of heads you get. What is the sample space?

{0, 1, 2, 3}

$X : \Omega \mapsto \mathbb{R}$

- Let Y be the number of tosses until you get the first head. What is the sample space?

{1, 2, 3, 4, ...}

Discrete random variables: A random variable is called discrete if the sample space is finite (e.g.: X above) or countably infinite (e.g.: Y above).
Counterexample: weight of a randomly selected person, waiting time until the next bus, etc. These are called *continuous* random variables.

Probability mass function

To fully define a random variable, we need to specify its distribution. In the discrete case, this is done by **assigning probability to each outcome in the sample space**.

- Let X be the outcome of rolling one fair die.
  - Sample space: $\{1, 2, 3, 4, 5, 6\}$
  - Corresponding probabilities: $P(X=1) = \frac{1}{6}$, ..., $P(X=6) = \frac{1}{6}$

notation: $P(1) = \frac{1}{6}$ or $P(x) = \frac{1}{6}$, $x = 1, 2, \cdots 6$

The function $P(x), x = 1, \ldots, 6$ is called the probability mass function for the discrete random variable X.

A valid probability mass function must satisfy the following basic conditions:

i. $P(x) \geq 0$ for all x in the sample space. EX: cannot assign $P(X=1) = -0.5$

ii. $\sum_{x \in sample\ space} P(x) = 1$. EX: $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$

Simulation

Expect: proportion of 1's to be close to $\frac{1}{6}$

We simulate the process of rolling one fair die to verify the probabilities, using https://www.random.org/dice/. What can we *expect* from the result?

- Rolling once: any outcome, by random chance
- Rolling 10 times: 2, 4, 1, 3, 1, 2, 3, 5, 5, 2 $\Rightarrow$ Proportion of 1: 0.2
- Rolling many, many times (1000? 10000?):

① There is no "rule" for how many simulations to run

② The larger the #, the more accurate the result

Goal: estimate $P(X=1)$

Simulation often requires **repeating** an experiment for **a large number of times**. The result is more trustworthy and can *better* estimate the truth because the summary of data (called a "statistic") has less variation.

Simulation

Simulation has limitations **when the sample space is very large** or **when some outcomes in the sample space has very low chance of occurring**.

- Imagine you have a possibly unfair die with 1000 side and you want to estimate the probability of rolling a one.
- If you don't repeat rolling the die enough times, the outcome of 1 might not occur at all.

Example 1

$$P(x) = \begin{cases} 0.25 & x = 0, 2 \\ 0.5 & x = 1 \\ 0 & \text{o.w.} \end{cases}$$

Sample space = $\{0, 1, 2\}$

Let X be the number of heads in two coin tosses. The coins are assumed fair. (model)

a. What is the probability mass function of X?

$P(X=0) = \underline{25\%}$ $\quad P(X=1) = \underline{50\%}$ $\quad P(X=2) = \underline{25\%}$

b. Verify that this probability mass function satisfy the two conditions.

1. all nonnegative ✓
2. $25\% + 50\% + 25\% = 1$ ✓

c. How can you simulate from this probability mass function to verify the probabilities empirically?

→ simulate two coins : T
T

→ # of head : 0

assume it's easy to generate an outcome according to a given p.m.f

$\begin{cases} H & \text{w.p. } 0.5 \\ T & \text{w.p. } 0.5 \end{cases}$

repeat many, many times ; record # of head

$\{0, 1, 0, \cdots\cdots, \}$

$P(X=0) = 0.25$: Check proportion of 0's in the result.

Example 2

→ repeat many, many times: {L, L, W, W, L, W ....}

(Example 1.3 from *Introduction to Probability*) Simulation is especially useful when the probabilities are hard to calculate mathematically. Suppose we roll a pair of fair dice. A gambler bets that, in 10 rolls, a pair of six would show up. What is his chance of winning? Describe how you would run a simulation to estimate his chance of winning.

win!
↑
↳ {1,2} ... .. {6,6}

first die: roll 10 times
second die: roll 10 times
sum: {5, 6, 5, 12, ...}
If w/in 10 times, there is a 12, record result as "win"
Else record as "lose"

(Can you calculate the exact probability mathematically?)

P(at least one pair of 6) = 1 − P(no pair of 6) $= 1 - \left(\frac{35}{36}\right)^{10}$

Example 3

(Example 1.4 from *Introduction to Probability*) You play a game with me as follows. A fair coin is tossed. If a head shows up, I give you a dollar. If a tail shows up, you give me a dollar.

a. Suppose we play this game 10 times. Let X be your net winning after 10 times. Can you design a simulation to study the distribution of X?

Estimate $P(X=-10)$

$P(X=-8)$

.....

Sample space : $\{-10, \cancel{-9}, -8, \cdots, \cancel{9}, 10\}$

b. Suppose you keep playing this game *until you win \$5*. Let Y be the length of the game you have to play. Can you design a simulation to study the distribution of Y?

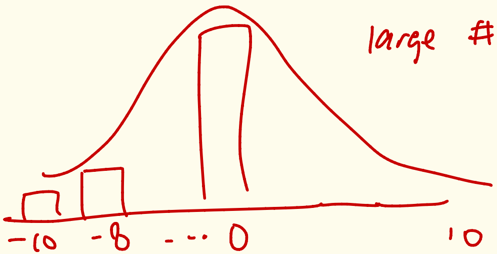
match : $\underbrace{+1,\ -1,\ -1,\ +1,\ \cdots}_{10 \text{ times}}$

Winning = sum of "match" $\Rightarrow$ 4

$\downarrow$

repeat many times

sequence of winning: $\underbrace{4,\ -2,\ 8,\ \cdots\cdots}_{\text{large \# of times}}$

↳ 

$\longrightarrow$ estimate of the distr. of X while winning < 5 <u>keep playing</u> ...

{ match : +1, -1, -1, ...

winning = sum (match) }

record length of game = 11

repeat many times ....

record length of game : $\underbrace{11,\ 225,\ \ldots\ldots,}_{\text{many times}}$