

## DISCRETE PROBABILITY MODELING AND SIMULATION

### 1. Reading

- Read section 2.3 from [1].
- Read sections 1.1,1.2,1.5.1,2.1 and 2.8 [2]

### 2. Learning objectives

- Familiarity with basic concepts in probability (events, probability distribution, conditioning, means, variances)
- Basics of python programming (arrays, Dataframes, plotting)
- The concept of Monte Carlo simulation

### 3. Statistical models

- **Models** are simplified representations of the world. There are many ways to represent models, but in science (and life), we often use **mathematical models**.
- The subject of this course is **regression models**. We will define this precisely later, but roughly speaking a regression model tells us how the distribution of a variable  $y$  (the response variable) is related to another variable  $x$  (the predictor).
- Examples: Predicting height based on age, predicting the probability to get disease given a mutation

### 4. Discrete random variables and distributions

- The rigorous mathematical theory for random variables is very useful, but requires certain machinery which is beyond the scope of these notes. Fortunately, we go a long way without such formalism. For our purposes, we can pretty much think of a random variable as any variable which we cannot predict prior to an observation of it, regardless of how much information we have. The classic example is the flip of a coin.
- The **sample space** is all the possible values that a random variable may take on. For the coin this would be heads or tails, for the roll of a dice  $1, 2, \dots, 6$  for the dice and the height of a person would be any positive number. Usually the outcomes are numbers, even if we use a number to represent a non-numerical quantity (e.g. someone's gender).
- In probability theory one distinguishes between outcomes and **events** – the latter are subsets of outcomes. For example, we might refer to the event that the roll of a die is greater than 2.
- We can characterize a random variable using a **probability model** or **probability distribution**, which maps a set of possible events to real numbers between 0 and 1.
- The **Bernoulli distribution** is probably the simplest probability distribution. It models a variable with binary outcome, for example the result of a YES/NO survey or a COVID test. The probability distribution can be written as a piecewise function

$$(1) \quad P(Y = y) = \begin{cases} q & y = 0 \\ (1 - q) & y = 1 \end{cases}$$

- These formulas make sense for any  $0 \leq q \leq 1$ . We say that  $q$  is a **parameter** of the distribution. In order to state that a Bernoulli distribution is a model for some random variable  $Y$ , we write

$$(2) \quad Y \sim \text{Bernoulli}(q).$$

- In general a probability distribution is a function that describes how likely certain outcomes are.
- In general, if  $Y$  is some random variable which could take ANY  $y$ , we define a probability distribution as a function from the space of all possible outcomes to an interval  $[0, 1]$ .
- For example, the space of outcomes  $S = \{\text{heads, tails}\}$  or  $S = \{1, 2, 3, 4, 5, 6\}$  (a dice).
- There are some rules for such functions.
  - $P(U) \leq 1$  for  $U \subset S$
  - $P(U) \geq 0$  for  $U \subset S$ . Here  $\subset$  means a subset of  $S$ , for example  $\{1, 2, 3\} \subset \{1, 2, 3, 4, 5, 6\}$ .
  - For a disjoint family of sets  $U_i$

$$\sum_i P(U_i) = P(\cup_i U_i)$$

Here  $\cup$  means “or”, for example,

$$P(\text{heads} \cup \text{tails})$$

is the probability that a coin is either heads or tails, which is of course one for any reasonable model.

**Example 1.** Suppose we flip two fair coins and let  $Y_A$  and  $Y_B$  denote the outcomes. The sample space is

$$S = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

For example, the event  $(0, 0)$  means that both genes are zero. If the coins are not biased, then each of the outcomes above should have the same probability. Thus each should have probability  $p$ . I'll use the notation

$$P(Y_A = 1, Y_B = 0) = P_{A,B}(1, 0) = p$$

We also know by additivity

$$P_{A,B}(0,0) + P_{A,B}(0,1) + P_{A,B}(1,0) + P_{A,B}(1,1) = 4p = 1 \implies p = \frac{1}{4}.$$

Another way to write this is

$$P_{A,B}(1,1) = P_A(1)P_B(1) = \frac{1}{2} \times \frac{1}{2}$$

**Example 2.** (Geometric distribution) Suppose we flip a fair coin until we see a heads. Let  $Y$  be the number of flips until we see a heads. This is example of a **geometric distribution**, which is the number of trials of independent, identically distributed (iid) Bernoulli random variables until we see  $k$  successes. In more mathematical notation, if

$$X_i \sim \text{Bernoulli}(q), \quad i = 1, 2, 3, \dots$$

then

$$Y = \min_{i \geq 0} \{i : X_i = 1\}$$

and we would say

$$Y \sim \text{Geometric}(p).$$

The sample space of  $Y$  is  $\{1, 2, \dots, \infty\}$ . What is the probability distribution?

$$\begin{aligned} P(Y = k) &= P(X_1 = 0, X_2 = 0, \dots, X_{k-1} = 0, X_k = 1) \\ &= P(X_1 = 0) \cdots P(X_{k-1} = 0)P(X_k = 1) \\ &= (1 - q)^{k-1}q \end{aligned}$$

This has the expected properties of  $Y$ . In particular, it decays as  $k$  increases and the decay is faster the larger  $q$  is.

- More generally, for a distribution with a particular name and set of parameters, we will write

$$\text{Variable} \sim \text{Distribution}(\text{parameters}).$$

We will sometimes use  $\theta$  to denote the parameters.

- A measurement of a random variable is a **sample** and **statistical inference** is the process of estimating the parameters  $\theta$  from a sample of a random variable.
- Consider the example of a survey: let's suppose we don't have information about every student in the college. Rather, a survey of five students from this class is conducted, finding 4 yeses and 1 no. What is our best prediction of the total fraction of students in the college who answered YES? What assumption do we make when we answer this question?
- **Probabilities as fraction vs. belief** There are two different ways we can interpret a statement like: The probability someone in this room is over 6 feet is 95%. Either it can be interpreted as a measure how likely it is to find someone in the room over 6 feet, or if we were to hypothetically generate random samples over and over what fraction of them would contain someone over 6 feet.

## 5. Python as a tool for statistical modeling

- When we generate samples using a computer we call them **simulations**. We will use python to perform simulations, and it is therefore important to have a basic understanding of the python language. It is assumed that you will go through the separate python tutorial notebook. For convenience, we will cover some basic tasks in this Notebook

**Example 3** (Flipping coins ). Let  $J$  denote a random variable representing the number of times a coin is flipped before two heads appear in a row. In the class python notebook I've written code that simulates  $J$ .

**Example 4** (Probability model from code). Write down the probability distribution for the output  $y$  of the following code

```
> def myRV():
>     y = 0
>     flip1 = np.random.choice([0,1],p=[0.5,0.5])
>     if flip1 ==0:
>         y = 10
>     else:
>         flip2 = np.random.choice([0,1],p=[0.5,0.5])
>         if flip2 ==0:
>             y = 2
>         else:
>             y =0
>     return y
```

First, we determine the sample space. The possible values of  $y$  are 10, 2 and 0. So  $S = \{0, 2, 10\}$ . What is the chance of each of these? If flip1 is 0 then we return 10 and this happens with probability  $1/2$ . If it is not zero, then we return 2 with probability  $1/2$ , thus, the probability we return 2 is  $1/4$  and same for 0. In summary

$$\begin{aligned} P(y = 10) &= \frac{1}{4} \\ P(y = 0) &= \frac{1}{2} \\ P(y = 2) &= \frac{1}{2} \end{aligned}$$

### 5.1. Monte Carlo Simulation.

- Often, we run many simulations of a model in order to say something about the distribution without performing any analytical calculations. We call these **Monte Carlo** simulations.
- Monte Carlo simulations make use of the fact that we can always conceptualize probabilities as fraction of things. That is, if we have  $n$  samples of a variable  $Y$  and we want to estimate  $P(Y = y)$ , then we can count the number for which  $Y = y$  – we denote this as  $n(Y = y)$ , and divide by the total number:  $P(Y = y) \approx n(Y = y)/n$ .
- Questions concerning how many samples we need to generate to obtain meaningful estimates from Monte Carlo simulations will be addressed later on.

**Example 5** (Example 4 cont.). In the python notebook we test the probabilities from Example 4.

**Example 6** (Verifying a formula). In the python notebook we verify the formula derived above for the probability distribution of a geometric random variable.

## 6. Independence and conditioning

- We start by considering the example of two variables which are not independent.

**Example 7** (Gene model). In this case, we need a model of both variables together (For example, this could be the model of whether someone has a mutation at two sites the genome):

$$P(Y_A, Y_B) = \begin{cases} 1/2 & \text{if } Y_A = 0 \text{ and } Y_B = 0 \\ 1/8 & \text{if } Y_A = 0 \text{ and } Y_B = 1 \\ 1/8 & \text{if } Y_A = 1 \text{ and } Y_B = 0 \\ 1/4 & \text{if } Y_A = 1 \text{ and } Y_B = 1 \end{cases}$$

The sample space is the same as Example 1, but we can check that  $Y_A$  and  $Y_B$  are no longer independent:

$$P(Y_A = 0, Y_B = 0) = \frac{1}{2}$$

on the other hand

$$\begin{aligned} P(Y_A = 0) &= P((0, 1) \text{ or } (0, 0)) = P(0, 1) + P(0, 0) = \frac{1}{8} + \frac{1}{2} = \frac{5}{8} \\ P(Y_B = 0) &= P((0, 0) \text{ or } (1, 0)) = P(0, 0) + P(1, 0) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8} \end{aligned}$$

and  $25/64 \approx 0.39 \neq 1/2$ .

- When we have two variables we call  $P(Y_A, Y_B)$  is an example of a **joint distribution**. It tells us the probabilities for observing *both* variables together, e.g. observing a person with both mutations. In general, if  $Y_1, \dots, Y_k$  are random variables we will use  $P(Y_1, \dots, Y_k)$  to denote their joint distribution.
- The joint distribution does not directly tell us the probabilities of observing e.g. someone with only one mutation. This can be obtained via **marginalization** which we say above; that is, summing over the other variable:

$$(3) \quad P(Y_A) = \sum_y P(Y_A, y) = P(Y_A, Y_B = 0) + P(Y_A, Y_B = 1)$$

where in the general the sum is taken over all possible outcomes for the second variable.  $\mathbb{P}(Y_1)$  is defined similarly.

- In the example above

$$(4) \quad Y_A \sim \text{Bernoulli}\left(\frac{5}{8}\right).$$

This is the distribution of  $Y_A$  absent any knowledge of  $Y_B$ .

### 6.1. Conditioning.

- What if we are interested in the chance that someone has a mutation in gene  $A$  and we know they do not have a mutation in gene  $B$ ? In this case, we introduce the **conditional probability**  $P(Y_A = 1 | Y_B = 0)$ . This is defined as the chance that gene  $A$  has a mutation in a person if we know there is no mutation at gene  $B$ . If we want to think about this in terms of population averages, it is the fraction of mutations in gene  $A$  among only those people without mutations in gene  $B$ .
- More general,  $P(X | Y = y)$  is the distribution of  $X$  if we know the value of  $Y = y$ .

- I'll use  $N$  to denote the number of individuals in a population with a given gene configuration and  $n$  the total population size. Interpreting probabilities as fraction,

$$\begin{aligned} P(Y_A = 1 | Y_B = 0) &= \frac{N(Y_A = 1, Y_B = 0)}{N(Y_B = 0)} = \frac{N(Y_A = 1, Y_B = 0)/n}{N(Y_B = 0)/n} \\ &= \frac{P(Y_A = 1, Y_B = 0)}{P(Y_B = 0)} \end{aligned}$$

**Example 8** (Example 7 cont.). Consider Example 7. We would have

$$P(Y_A = 1 | Y_B = 0) = \frac{P(1, 0)}{P(Y_B = 0)} = \frac{1/8}{5/8} = \frac{1}{5}$$

We can easily perform this computation using simulated data in python. In the python notebook we use Monte carlo simulations to show they are not independent.

**Example 9** (Example 4 cont.). In the code from Example 4 cont. we have the internal variables `flip1` and `flip2`. If  $Y$  is the output, what is the conditional distribution

$$Y | (\text{flip1} == 1)$$

- In general, we have

$$(5) \quad P(Y|X) = \frac{P(Y, X)}{P(X)}.$$

Notice that we can replace  $P(Y, X) = P(Y|X)P(X)$ , to obtain Baye's formula

$$(6) \quad P(Y|X) = \frac{P(Y, X)}{P(X)}.$$

- Two variables are said to be **independent** if  $P(Y|X) = P(Y)$  and  $P(X|Y) = P(X)$ .
- Can you see why  $X$  being independent of  $Y$  implies  $Y$  is independent of  $X$ ? Equation (7) is also true for events, for example, we will encounter things like

$$(7) \quad P(Y > z | X) = \frac{P(Y > z, X)}{P(X)}$$

#### References

- [1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning (python version)*, volume 112. Springer, 2013.
- [2] John Tabak. *Probability and statistics: The science of uncertainty*. Infobase Publishing, 2014.