DS-GA 1002 Lecture notes 1 September 14, 2015

Probability

1 Introduction

Probability theory is a tool that allows us to reason mathematically about uncertainty. In this notes we will learn how to build probabilistic models to incorporate the information we have about uncertain variables of interest in a principled way.

2 Definition of probability space

In order to reason probabilistically about an uncertain phenomenon (the result of rolling a die, tomorrow's weather, the result of an NBA game, ...) we characterize it as an **experiment** with several (possibly infinite) mutually exclusive **outcomes**. Probability theory relies on set theory to model the uncertainty in the experiment (for those of you that might need a review, a separate handout provides some background on set theory). The possible outcomes of the experiment are grouped into sets called **events** which are assigned a measure according to how likely they are. More formally, the experiment is characterized by constructing a **probability space**.

Definition 2.1 (Probability space). *A probability space is a triple* $(\Omega, \mathcal{F}, \mathrm{P})$ *consisting of:*

- *A **sample space** Ω, which contains all possible outcomes of the experiment.*
- *A set of events $\mathcal{F}$, which must be a **σ-algebra** (see Definition 2.2 below).*
- *A **probability measure** P that assigns probabilities to the events in $\mathcal{F}$.*

The term σ-algebra is used in measure theory to denote a collection of sets that satisfy certain conditions listed below. Don't be too intimidated by it. It is just a sophisticated way of stating that if we assign a probability to certain events (for example *it will rain tomorrow* or *it will snow tomorrow*) we also need to assign a probability to their complements (i.e. *it will not rain tomorrow* or *it will not snow tomorrow*) and to their union (*it will rain or snow tomorrow*).

Definition 2.2 (σ-algebra). *A σ-algebra $\mathcal{F}$ is a collection of sets in Ω such that:*

1. *If a set $S \in \mathcal{F}$ then $S^c \in \mathcal{F}$.*

2. *If the sets $S_1, S_2 \in \mathcal{F}$, then $S_1 \cup S_2 \in \mathcal{F}$. This also holds for infinite sequences; if $S_1, S_2, \ldots \in \mathcal{F}$ then $\cup_{i=1}^{\infty} S_i \in \mathcal{F}$.*

3. *$\Omega \in \mathcal{F}$.*

The role of the probability measure P is to quantify how likely we are to encounter an outcome from each of the events in the σ-algebra.

Definition 2.3 (Probability measure)**.** *A probability measure is a function defined over the sets in a σ-algebra $\mathcal{F}$ such that:*

1. $\mathrm{P}(S) \geq 0$ *for any event* $S \in \mathcal{F}$.

2. *If the sets $S_1, S_2, \ldots, S_n \in \mathcal{F}$ are disjoint (i.e. $S_i \cap S_j = \emptyset$ for $i \neq j$) then*

$$\mathrm{P}\left(\cup_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1}$$

Similarly, for a countably infinite sequence of disjoint sets $S_1, S_2, \ldots \in \mathcal{F}$

$$\mathrm{P}\left(\lim_{n\to\infty} \cup_{i=1}^{n} S_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{2}$$

3. $\mathrm{P}(\Omega) = 1$.

The two first axioms capture the intuitive idea that the probability of an event is a measure such as mass (or length or volume): just like the mass of any object is nonnegative and the total mass of several distinct objects is the sum of their masses, the probability of any event is nonnegative and the probability of the union of several disjoint objects is the sum of their probabilities. However, in contrast to mass, the amount of probability in an experiment cannot be unbounded. If it is highly likely that it will rain tomorrow, then it cannot be also very likely that it will *not* rain. More formally, if the probability of an event S is large, then the probability of its complement S^c must be small. This is captured by the third axiom, which normalizes the probability measure (and implies that $\mathrm{P}(S^c) = 1 - \mathrm{P}(S)$).

To simplify notation, we write the probability of the intersection of several events in the following form

$$\mathrm{P}(A, B, C) := \mathrm{P}(A \cap B \cap C). \tag{3}$$

We now illustrate the above definitions with an example.

Example 2.4 (Probability space for a basketball game)**.** The Cleveland Cavaliers are playing the Golden State Warriors tomorrow and we want to model the game probabilistically. In order to do this, we construct the following probability space.

- **Sample space**:
  We choose to represent the outcome of the game by its final score.

$$\Omega = \{\text{Cavs } 1 - \text{Warriors } 0, \text{Cavs } 0 - \text{Warriors } 1, \ldots, \text{Cavs } 101 - \text{Warriors } 97, \ldots\}. \tag{4}$$

- **σ-algebra**:
  We are only interested in what team wins, so

$$\mathcal{F} = \{\text{Cavs win}, \text{Warriors win}, \text{Cavs or Warriors win}, \emptyset\}. \tag{5}$$

- **Probability measure**:
  We believe that the two teams are equally likely to win:

$$\mathrm{P}(\text{Cavs win}) = \frac{1}{2}, \quad \mathrm{P}(\text{Cavs win}) = \frac{1}{2}, \quad \mathrm{P}(\text{Cavs or Warriors win}) = 1, \quad \mathrm{P}(\emptyset) = 0. \tag{6}$$

This example reveals some interesting aspects of probability spaces.

- The probability measure does not necessarily assign a probability to individual outcomes, but rather to events (which are *sets* of outcomes). This distinction is important for continuous sample spaces, as we discuss in Section 5.2.1.
- The σ-algebra $\mathcal{F}$ determines with what granularity we analyze the experiment. Under the σ-algebra chosen in Example 2.4 we do not have access to events such as *the Cavs score less than 100 points* or *the Warriors score 10 points more than the Cavs.*

Let us list some important consequences of Definition 2.3:

$$\mathrm{P}(\emptyset) = 0, \tag{7}$$

$$A \subseteq B \quad \text{implies} \quad \mathrm{P}(A) \leq \mathrm{P}(B), \tag{8}$$

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{9}$$

Finally, we record a simple yet extremely useful result that allows to bound the probability of the union of a collection of events by the sum of their individual probabilities.

Theorem 2.5 (Union bound). *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $S_1, S_2, \ldots$ a collection of events in $\mathcal{F}$. Then*

$$\mathrm{P}\left(\cup_i S_i\right) \leq \sum_i \mathrm{P}\left(S_i\right). \tag{10}$$

Proof. Let us define the sets:

$$\tilde{S}_i = S_i \cap \cap_{j=1}^{i-1} S_j^c. \tag{11}$$

It is straightforward to show by induction that $\cup_{j=1}^n S_j = \cup_{j=1}^n \tilde{S}_j$ for any n, so $\cup_i S_i = \cup_i \tilde{S}_i$. The sets $\tilde{S}_1$, $\tilde{S}_2$, ... are disjoint by construction, so

$$\mathrm{P}\left(\cup_i S_i\right) = \mathrm{P}\left(\cup_i \tilde{S}_i\right) = \sum_i \mathrm{P}\left(\tilde{S}_i\right) \quad \text{by Axiom 2 in Definition 2.3} \tag{12}$$

$$\leq \sum_i \mathrm{P}\left(S_i\right) \quad \text{because } \tilde{S}_i \subseteq S_i. \tag{13}$$

□

3 Conditional probability

The concept of conditional probability allows us to update probabilistic models if some information is revealed. Given a probabilistic space $(\Omega, \mathcal{F}, \mathrm{P})$ we might want to assume that a certain event $S \in \mathcal{F}$ with nonzero probability (i.e. $\mathrm{P}(S) > 0$) has occurred. In other words, we know that the outcome of the experiment belongs to S, *but that is all we know.* In order to incorporate this information, we define the **conditional probability** of an event $S' \in \mathcal{F}$ **given** S as

$$\mathrm{P}\left(S' | S\right) := \frac{\mathrm{P}\left(S' \cap S\right)}{\mathrm{P}\left(S\right)}. \tag{14}$$

This definition is rather intuitive: We are assuming that S occurs, so if the outcome is in S' then it must belong to $S' \cap S$. However, we cannot just take the probability of the intersection because the sample space has been reduced to S. Therefore we normalize by the probability of S. As a sanity check, we have $\mathrm{P}(S|S) = 1$ and $\mathrm{P}(S'|S) = 0$ if S and S' are disjoint.

The conditional probability $\mathrm{P}(\cdot|S)$ is a valid probability measure in the probability space $(S, \mathcal{F}_S, \mathrm{P}(\cdot|S))$, where $\mathcal{F}_S$ is a σ-algebra that contains the intersection of S and the sets in $\mathcal{F}$.

Example 3.1 (Flights and rain)**.** JFK airport hires you to estimate how the punctuality of flight arrivals is affected by the weather. You begin by defining a probability space for which the sample space is

$$\Omega = \{\text{late and rain}, \text{late and no rain}, \text{on time and rain}, \text{on time and no rain}\} \tag{15}$$

and the σ-algebra is 2^Ω the power set of Ω. From data of past flights you determine that a reasonable estimate for a probability measure is

$$\mathrm{P}\left(\text{late}, \text{no rain}\right) = \frac{2}{20}, \quad \mathrm{P}\left(\text{on time}, \text{no rain}\right) = \frac{14}{20}, \tag{16}$$

$$\mathrm{P}\left(\text{late}, \text{rain}\right) = \frac{3}{20}, \quad \mathrm{P}\left(\text{on time}, \text{rain}\right) = \frac{1}{20}. \tag{17}$$

The airport is interested in the probability of a flight being late if it rains, so you define a new probability space conditioning on the event *rain*. The sample space is the set of all outcomes such that *rain* occurred, the σ-algebra is the power set of $\{\text{on time}, \text{late}\}$ and the probability measure is $\mathrm{P}\left(\cdot|\text{rain}\right)$. In particular,

$$\mathrm{P}\left(\text{late}|\text{rain}\right) = \frac{\mathrm{P}\left(\text{late}, \text{rain}\right)}{\mathrm{P}\left(\text{rain}\right)} = \frac{3/20}{3/20 + 1/20} = \frac{3}{4} \tag{18}$$

and similarly $\mathrm{P}\left(\text{late}|\text{no rain}\right) = 1/8$.

Conditional probabilities are useful to compute the intersection of several events in a structured way. By definition, we can express the probability of the intersection of two events $A, B \in \mathcal{F}$ as follows,

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A\right) \mathrm{P}\left(B|A\right) = \mathrm{P}\left(B\right) \mathrm{P}\left(A|B\right). \tag{19}$$

In this formula $\mathrm{P}\left(A\right)$ is known as the **prior** probability of A, as it captures the information we have about A before anything else is revealed. Analogously, $\mathrm{P}\left(A|B\right)$ is known as the **posterior** probability. Generalizing (19) to a sequence of events gives the *chain rule*, which allows to express the probability of the intersection of multiple events in terms of conditional probabilities. We omit the proof, which is a straightforward application of induction.

Theorem 3.2 (Chain rule)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $S_1, S_2, \ldots$ a collection of events in $\mathcal{F}$,*

$$\mathrm{P}\left(\cap_i S_i\right) = \mathrm{P}\left(S_1\right) \mathrm{P}\left(S_2|S_1\right) \mathrm{P}\left(S_3|S_1 \cap S_2\right) \cdots = \prod_i \mathrm{P}\left(S_i|\cap_{j=1}^{i-1} S_j\right). \tag{20}$$

Sometimes, estimating the probability of a certain event directly may be more challenging than estimating its probability conditioned on other simpler events that cover the whole sampling space. The Law of Total Probability allows us to pool these conditional probabilities together, weighing them by the probability of the individual events, to compute the probability of the event of interest.

Theorem 3.3 (Law of Total Probability)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $P_1, P_2, \ldots \in \mathcal{F}$ *a partition of* Ω *(i.e. they are disjoint and* $\Omega = \cup_i P_i$*). For any set* $S \in \mathcal{F}$

$$\mathrm{P}(S) = \sum_i \mathrm{P}(S \cap P_i) = \sum_i \mathrm{P}(P_i)\,\mathrm{P}(S|P_i)\,. \tag{21}$$

Proof. This is an immediate consequence of the chain rule and Axiom 2 in Definition 2.3, since $S = \cup_i S \cap P_i$ and the sets $S \cap P_i$ are disjoint. □

Example 3.4 (Aunt visit)**.** Your aunt is arriving at JFK tomorrow and you would like to know how likely it is for her flight to be on time. From Example 3.1, you recall that

$$\mathrm{P}(\text{late}|\text{rain}) = 0.75, \quad \mathrm{P}(\text{late}|\text{no rain}) = 0.125. \tag{22}$$

After checking out a weather website, you determine that $\mathrm{P}(\text{rain}) = 0.2$.

Now, how can we integrate all of this information? The events *rain* and *no rain* are disjoint and cover the whole sample space, so they form a partition. We can consequently apply the Law of Total Probability to determine

$$\mathrm{P}(\text{late}) = \mathrm{P}(\text{late}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain}) \tag{23}$$

$$= 0.75 \cdot 0.2 + 0.125 \cdot 0.8 = 0.25. \tag{24}$$

So the probability that your aunt's plane is late is 1/4.

It is crucial to realize that in general $\mathrm{P}(A|B) \neq \mathrm{P}(A|B)$: if you are a professional tennis player you probably started playing when you were young, but having played from a young age sure doesn't imply a high probability of playing professionally. The reason is that the prior probabilities will not be the same: in the example, the probability of being a professional is much lower than the probability of playing tennis as a kid. However, it is possible to *invert* conditional probabilities, i.e. find $\mathrm{P}(A|B)$ from $\mathrm{P}(B|A)$ by reweighing it adequately. The ensuing formula is called Bayes' Rule.

Theorem 3.5 (Bayes' Rule)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $P_1, P_2, \ldots \in \mathcal{F}$ *a partition of* Ω *(i.e. they are disjoint and* $\Omega = \cup_i P_i$*). For any set* $S \in \mathcal{F}$

$$\mathrm{P}\left(P_i|S\right) = \frac{\mathrm{P}\left(S|P_i\right)\mathrm{P}\left(P_i\right)}{\sum_j \mathrm{P}\left(S|P_j\right)\mathrm{P}\left(P_j\right)}. \tag{25}$$

Proof. The theorem follows from the definition of conditional probability and the Law of Total Probability (3.3). □

Example 3.6 (Aunt visit (continued))**.** You explain the probabilistic model described in Example 3.4 to your cousin Marvin who lives in California. A day later, you tell him that your aunt arrived late but you don't mention whether it rained or not. After he hangs up, Marvin wants to figure out the probability that it rained. Recall that the probability of rain was 0.2, but since your aunt arrived late he should update the estimate. Applying Bayes' Rule:

$$\begin{align} \mathrm{P}\left(\text{rain}|\text{late}\right) &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right) + \mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} \tag{26} \\ &= \frac{0.75 \cdot 0.2}{0.75 \cdot 0.2 + 0.125 \cdot 0.8} = 0.6. \tag{27} \end{align}$$

As expected, the probability that it rained in New York is significantly higher now that Marvin knows that your aunt was late.

---

4 Independence

As we saw in the previous section, conditional probabilities quantify the extent to which the knowledge of the occurrence of a certain event affects how likely it is for another event to happen. In some cases, there is no difference at all: the events are **independent**. More formally, events A and B are independent if

$$\mathrm{P}\left(A|B\right) = \mathrm{P}\left(A\right). \tag{28}$$

This definition is not valid if $\mathrm{P}\left(B\right) = 0$, so below we give a condition that covers this case but that is otherwise equivalent.

Definition 4.1 (Independence)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. Two sets* $A, B \in \mathcal{F}$ *are independent if*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B). \tag{29}$$

A collection of sets $S_1, S_2, \ldots \in \mathcal{F}$ *is independent if*

$$\mathrm{P}(\cap_i S_i) = \prod_i \mathrm{P}(S_i). \tag{30}$$

Similarly, we can define **conditional independence** between two sets given a third set.

Definition 4.2 (Conditional independence)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. Two sets* $A, B \in \mathcal{F}$ *are conditionally independent given a third set* $C \in \mathcal{F}$ *if*

$$\mathrm{P}(A \cap B | C) = \mathrm{P}(A|C)\,\mathrm{P}(B|C). \tag{31}$$

A collection of sets $S_1, S_2, \ldots \in \mathcal{F}$ *is conditionally independent given another set* $S \in \mathcal{F}$ *if*

$$\mathrm{P}(\cap_i S_i | S) = \prod_i \mathrm{P}(S_i | S). \tag{32}$$

Interestingly, as we will see in the examples below, independence does not imply conditional independence or vice versa.

Example 4.3 (Conditional independence does not imply independence)**.** Your cousin Marvin from Exercise 3.6 always complains about taxis in New York. From his many visits to JFK he has calculated that

$$\mathrm{P}(\text{taxi}|\text{rain}) = 0.1, \quad \mathrm{P}(\text{taxi}|\text{no rain}) = 0.6, \tag{33}$$

where *taxi* denotes the event of finding a free taxi after picking up your luggage. Given the event *rain*, it is reasonable to model the events *plane arrived late* and *taxi* as conditionally independent,

$$\mathrm{P}(\text{taxi}, \text{late}|\text{rain}) = \mathrm{P}(\text{taxi}|\text{rain})\,\mathrm{P}(\text{late}|\text{rain}), \tag{34}$$

$$\mathrm{P}(\text{taxi}, \text{late}|\text{no rain}) = \mathrm{P}(\text{taxi}|\text{no rain})\,\mathrm{P}(\text{late}|\text{no rain}). \tag{35}$$

The logic behind this is that the availability of taxis after picking up your luggage depends on whether it's raining or not, but not on whether the plane is late or not (let us assume that the availability is constant throughout the day). Does this assumption imply that the events are independent?

If they were independent, then knowing that your aunt was late would give no information to Marvin about taxi availability. However,

$$\begin{aligned}
\mathrm{P}(\text{taxi}) &= \mathrm{P}(\text{taxi}, \text{rain}) + \mathrm{P}(\text{taxi}, \text{no rain}) \quad \text{(by the Law of Total Probability)} && (36)\\
&= \mathrm{P}(\text{taxi}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{taxi}|\text{no rain})\,\mathrm{P}(\text{no rain}) && (37)\\
&= 0.1 \cdot 0.2 + 0.6 \cdot 0.8 = 0.5, && (38)\\
\mathrm{P}(\text{taxi}|\text{late}) &= \frac{\mathrm{P}(\text{taxi}, \text{late}, \text{rain}) + \mathrm{P}(\text{taxi}, \text{late}, \text{no rain})}{\mathrm{P}(\text{late})} \quad \text{(by the Law of Total Probability)} && (39)\\
&= \frac{\mathrm{P}(\text{taxi}|\text{rain})\,\mathrm{P}(\text{late}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{taxi}|\text{no rain})\,\mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain})}{\mathrm{P}(\text{late})} && (40)\\
&= \frac{0.1 \cdot 0.75 \cdot 0.2 + 0.6 \cdot 0.125 \cdot 0.8}{0.25} = 0.3. && (41)
\end{aligned}$$

$\mathrm{P}(\text{taxi}) > \mathrm{P}(\text{taxi}|\text{late})$ so the events are *not* independent. This makes sense, since if the airplane is late, it is more probable that it is raining, which makes taxis more difficult to find.

Example 4.4 (Independence does not imply conditional independence)**.** After looking at your probabilistic model from Example 3.1 your contact at JFK points out that delays are often caused by mechanical problems in the airplanes. You look at the data and determine that

$$\mathrm{P}(\text{problem}) = \mathrm{P}(\text{problem}|\text{rain}) = \mathrm{P}(\text{problem}|\text{no rain}) = 0.1, \tag{42}$$

so the events *mechanical problem* and *rain in NYC* are independent, which makes intuitive sense. After some more data analysis, you estimate

$$\mathrm{P}(\text{late}|\text{problem}) = 0.7, \quad \mathrm{P}(\text{late}|\text{no problem}) = 0.2, \quad \mathrm{P}(\text{late}|\text{no rain}, \text{problem}) = 0.5. \tag{43}$$

The next time you are waiting for Marvin at JFK, you start wondering about the probability of his plane having had some mechanical problem. Without any further information, this probability is 0.1. It is a sunny day in New York, but this is of no help because according to the data (and common sense) the events *problem* and *rain* are independent.

Suddenly they announce that Marvin's plane is late. Now, what is the probability that his plane had a mechanical problem? At first thought you might apply Bayes' Rule to compute

$\mathrm{P}(\text{problem}|\text{late}) = 0.28$ as in Example 3.6. However, you are not using the fact that it is sunny. This means that the rain was not responsible for the delay, so intuitively a mechanical problem should be more likely. Indeed,

$$\begin{aligned}
\mathrm{P}(\text{problem}|\text{late}, \text{no rain}) &= \frac{\mathrm{P}(\text{late}, \text{no rain}, \text{problem})}{\mathrm{P}(\text{late}, \text{no rain})} \tag{44}\\
&= \frac{\mathrm{P}(\text{late}|\text{no rain}, \text{problem})\,\mathrm{P}(\text{no rain})\,\mathrm{P}(\text{problem})}{\mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain})} \quad \text{(by the Chain Rule)} \tag{45}\\
&= \frac{0.5 \cdot 0.1}{0.125} = 0.4. \tag{46}
\end{aligned}$$

Since $\mathrm{P}(\text{problem}|\text{late}, \text{no rain}) \neq \mathrm{P}(\text{problem}|\text{late})$ the events *mechanical problem* and *rain in NYC* are *not* conditionally independent given the event *plane is late*.

5 Random variables

Random variables are a fundamental tool in probabilistic modeling. They allow us to characterize numerical quantities that are *uncertain*: the temperature in New York tomorrow, the speed of an airplane at a certain time, the number of goals that will be scored by Messi next year... Reasoning about such quantities probabilistically allows us to structure the information we have about them in a principled way.

Formally, we define random variables as functions of the outcomes in a probability space.

Definition 5.1 (Random variable). *Given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, a random variable X is a function from the sample space Ω to the real numbers $\mathbb{R}$. Once the outcome $\omega \in \Omega$ of the experiment is revealed, the corresponding $X(\omega)$ is known as the* ***realization*** *of the random variable.*

Remark 5.2 (Rigorous definition). *If we want to be completely rigorous, Definition 5.1 is missing some details. Consider two sample spaces Ω_1 and Ω_2, and a σ-algebra $\mathcal{F}_2$ of sets in Ω_2. Then, for X to be a random variable, there must exist a σ-algebra $\mathcal{F}_1$ in Ω_1 such that for any set S in $\mathcal{F}_2$ the inverse image of S, defined by*

$$X^{-1}(S) := \{\omega \mid X(\omega) \in S\}, \tag{47}$$

belongs to $\mathcal{F}_1$. Usually, we take Ω_2 to be the reals $\mathbb{R}$ and $\mathcal{F}_2$ to be the Borel σ-algebra, which is defined as the smallest σ-algebra defined on the reals that contains all open intervals *(amazingly, it is possible to construct sets of real numbers that do not belong to this σ-algebra). In any case, for the purpose of this course, Definition 5.1 is sufficient. If you are interested in learning more about the formal foundations of probability we encourage you to take a course in measure theory and advanced probability theory.*

You should **not** think of a random variable as having a fixed numerical value, even if we already know the outcome of the phenomenon of interest: that is precisely what a realization of the random variable represents. In contrast, the random variable captures the uncertainty in our probabilistic modeling. In order to stress the difference between random variables and their realizations, we denote the former with uppercase letters $(X, Y, \ldots)$ and the latter with lowercase letters $(x, y, \ldots)$.

If we have access to the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ in which the random variable is defined then it is straightforward to compute the probability of a random variable X belonging to a certain set $S \subseteq \mathbb{R}$: it is the probability of the event that comprises all outcomes in Ω which are mapped to S under X,

$$\mathrm{P}(X \in S) = \mathrm{P}(\{\omega \mid X(\omega) \in S\}). \tag{48}$$

However, we almost never model the probability space directly, since this requires estimating the probability of any possible event. Instead, there are other ways to specify random variables, which we will describe in Sections 5.1 and 5.2.1, that imply that a valid underlying probability space exists. This probability space is useful mainly from a theoretical point of view; it ensures that the whole framework is mathematically sound, but you don't really have to worry about it.

There are two main kinds of random variables:

- **Discrete** random variables take values on a finite or countably infinite subset of $\mathbb{R}$ such as the integers.
- **Continuous** random variables take values over the real line $\mathbb{R}$.

5.1 Discrete random variables

Discrete random variables are numerical quantities that take either finite or countably infinite values: the outcome of the roll of a die, the score of a team in a basketball game, etc.

5.1.1 Probability mass function

To specify a discrete random variable it is enough to determine the probability of each possible value that it can take.

Definition 5.3 (Probability mass function)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $X : \Omega \to \mathbb{Z}$ *a random variable. The* ***probability mass function*** *(pmf) of* X *is defined as*

$$p_X(x) := \mathrm{P}(\{\omega \mid X(\omega) = x\}). \tag{49}$$

In words, $p_X(x)$ *is the probability that* X *equals* x*.*

We usually say that a random variable is **distributed** according to a certain pmf.

If the discrete domain of X is denoted by D, then the triplet $\left(D, 2^D, p_X\right)$ is a valid probability space (recall that 2^D is the power set of D). In particular, p_x is a valid probability measure which satisfies

$$p_X(x) \geq 0 \quad \text{for any } x \in D, \tag{50}$$

$$\sum_{x \in D} p_X(x) = 1. \tag{51}$$

The converse is also true, if a function defined on a countable subset D of the reals is nonnegative and it adds up to one, then it may be interpreted as the pmf of a random variable. In fact, this is usually how we define the random variables that we work with.

To compute the probability that a random variable X is in a certain set S we take the sum of the pmf over all the values contained in S:

$$\mathrm{P}(X \in S) = \sum_{x \in S} p_X(x). \tag{52}$$

5.1.2 Important discrete random variables

In this section we describe several discrete random variables that are very popular in probabilistic modeling.

Bernouilli random variables are used to model experiments that have two possible outcomes. By convention we usually represent an outcome by 0 and the other outcome by 1. A canonical example is flipping a biased coin, such that the probability of obtaining heads is p. If we encode heads as 1 and tails as 0, then the result of the coin flip corresponds to a Bernouilli random variable with parameter p.

Definition 5.4 (Bernouilli)**.** *The pmf of a Bernouilli random variable with parameter* $p \in [0, 1]$ *is given by*

$$p_X(0) = 1 - p, \tag{53}$$

$$p_X(1) = p. \tag{54}$$

A special kind of Bernouilli random variable is the indicator random variable of an event. This random variable is particularly useful in proofs.

Definition 5.5 (Indicator)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. The indicator random variable of an event* $S \in \mathcal{F}$ *is defined as*

$$1_S(\omega) = \begin{cases} 1, & \text{if } \omega \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{55}$$

An indicator random variable is Bernouilli with parameter $\mathrm{P}(S)$*.*

Imagine that we take a biased coin and flip it until we obtain heads. The number of flips that we need can be modeled as a **geometric** random variable, as long as the flips are independent (this follows directly from (30)).

Definition 5.6 (Geometric)**.** *The pmf of a geometric random variable with parameter* p *is given by*

$$p_X(k) = (1-p)^{k-1}\, p, \quad k = 1, 2, \ldots \tag{56}$$

A geometric random variable models the number of times that we need to repeat an experiment with a binary outcome to obtain a certain result. In contrast, a **binomial** random variable models the number of times we obtain a certain result over a fixed number of repetitions.

Definition 5.7 (Binomial)**.** *The pmf of a binomial random variable with parameters* n *and* p *is given by*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad k = 0, 1, 2, \ldots, n. \tag{57}$$

Recall that

$$\binom{n}{k} := \frac{n!}{k!\,(n-k)!} \tag{58}$$

is the binomial coefficient, which is the number of ways that we can choose k elements from a group of n elements. The following lemma shows that in terms of coin flips, a binomial models the number of heads we obtain if we perform n independent flips.

Lemma 5.8. *The distribution of the number of heads obtained in* n *independent flips using a coin with a bias of* p *is binomial.*

Proof. We need to prove that if X is binomial with parameters n and p then

$$\mathrm{P}\left(k \text{ heads}\right) = p_X\left(k\right) = \binom{n}{k} p^k \left(1-p\right)^{(n-k)}. \tag{59}$$

Any fixed flip sequence with k heads has probability $p^k \left(1-p\right)^{(n-k)}$ by (30) because the flips are independent. The event k *heads* is the union of all possible sequences of this form. These events are disjoint and there are $\binom{n}{k}$ of them, so the conclusion follows from (1). □

We motivate the definition of the **Poisson** random variable using an example.

Example 5.9 (Emails per day)**.** Imagine that you want to model the number of emails that you receive each day. You make the following assumptions:

- Each email is sent independently from every other email.
- Each email is sent at any point of the day with the same probability that only depends on a certain *rate* λ measured in emails per day (note that this does **not** mean that you receive the λ emails per day; otherwise there is no need for a probabilistic model).

If we discretize the day into n intervals for n large enough that two emails never arrive at the same time, then we can model the number of emails by a binomial with parameters n and $p = \lambda/n$: there are n slots and an email arrives in each slot with probability λ/n independently from any other email. Note that we have normalized the rate λ because its units are emails per day.

Now, we can compute the distribution of the number of emails when the grid is arbitrarily small, i.e. when $n \to \infty$:

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{P}\left(k \text{ emails in } n \text{ slots}\right) &= \lim_{n\to\infty} \binom{n}{k} p^k \left(1-p\right)^{(n-k)} && (60)\\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{(n-k)} && (61)\\
&= \lim_{n\to\infty} \frac{n!\,\lambda^k}{k!\,(n-k)!\,(n-\lambda)^k} \left(1-\frac{\lambda}{n}\right)^n && (62)\\
&= \frac{\lambda^k\, e^{-\lambda}}{k!}. && (63)
\end{aligned}$$

The last step follows from the following lemma proved in Section A of the appendix.

Lemma 5.10.

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,(n-\lambda)^k}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{64}$$

As illustrated in the example, Poisson random variables model the number of events in an interval if the events occur independently with a certain rate.

Definition 5.11 (Poisson)**.** *The pmf of a Poisson random variable with parameter λ is given by*

$$p_X(k) = \frac{\lambda^k\, e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots \tag{65}$$

5.2 Continuous random variables

We tend to think of many physical quantities as being continuous: temperature, duration, speed, weight, etc. In order to model such quantities probabilistically we could discretize their domain and use discrete random variables. However, we may not want our conclusions to depend on how we choose the discretization grid. Constructing a continuous model allows to obtain insights that are valid for *sufficiently fine* grids without worrying about the discretization parameters.

Precisely because continuous domains model the limit when discrete outcomes have an arbitrarily fine granularity, we **cannot** characterize the probabilistic behavior of a continuous random variable by just setting values for the probability of X being equal to individual outcomes, as we do for discrete random variables. In other words, continuous random variables *cannot* have a probability mass function assigning probabilities to specific outcomes. Intuitively, if we have uncountable disjoint outcomes with nonzero probability, then the probability of their union is infinite, which contradicts Definition 2.3 as there cannot be an event with probability greater than one.

More rigorously, it turns out that we cannot define a valid probability measure on the power set of $\mathbb{R}$ (justifying this requires measure theory and is beyond the scope of these notes). Instead, we only consider events that are composed of *unions of intervals*. Such events form a σ-algebra called the Borel σ-algebra. This σ-algebra is granular enough to represent any set that you might be interested in (try thinking of a set that cannot be expressed as a countable union of intervals), while allowing for valid probability measures to be defined on it.

5.2.1 Cumulative distribution function and probability density function

To specify a random variable on the Borel σ-algebra it suffices to determine the probability of the random variable belonging to all intervals of the form $(-\infty, x)$ for $x \in \mathbb{R}$.

Definition 5.12 (Cumulative distribution function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. The* ***cumulative distribution function*** *(cdf) of X is defined as*

$$F_X(x) := \mathrm{P}(X \leq x). \tag{66}$$

In words, $F_X(x)$ is the probability of X being smaller than x.

Note that the cumulative distribution function can be defined for *both continuous and discrete* random variables.

The following lemma describes some basic properties of the cdf. You can find the proof in Section B of the appendix.

Lemma 5.13 (Properties of the cdf)**.** *For any continuous random variable X*

$$\lim_{x \to -\infty} F_X(x) = 0, \tag{67}$$

$$\lim_{x \to \infty} F_X(x) = 1, \tag{68}$$

$$F_X(b) \geq F_X(a) \quad \textit{if } b > a, \quad \textit{i.e. } F_X \textit{ is nondecreasing.} \tag{69}$$

To see why the cdf completely determines a random variable recall that we are only considering sets that can be expressed as unions of intervals. The probability of a random variable X belonging to an interval $(a, b]$ is given by

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = F_X(b) - F_X(a). \tag{70}$$

Remark 5.14. *Since individual points have zero probability, for any continuous random variable X*

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X < b) = \mathrm{P}(a \leq X < b). \tag{71}$$

Now, to find the probability of X belonging to any particular set, we only need to decompose it into disjoint intervals and apply (71).

If the cdf of a continuous random variable is differentiable, its derivative can be interpreted as a **probability density function**.

Definition 5.15 (Probability density function)**.** *Let $X : \Omega \to \mathbb{Z}$ be a random variable with cdf F_X. If F_X is differentiable then the probability density function or* ***pdf*** *of X is defined as*

$$f_X(x) := \frac{dF_X(x)}{d\,x}. \tag{72}$$

Intuitively, for an interval of width Δ, $f_X(x)\,\Delta$ is the probability of X being in the interval Δ as $\Delta \to 0$. From the fundamental theorem of calculus it follows that the probability of a random variable X belonging to an interval is given by

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,\mathrm{d}x. \tag{73}$$

Since we are considering sets in the Borel σ-algebra, which can be decomposed into unions of intervals, it follows that we can obtain the probability of X belonging to any set S by integrating its pdf over S

$$\mathrm{P}(X \in S) = \int_S f_X(x)\,\mathrm{d}x. \tag{74}$$

In particular,

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1. \tag{75}$$

It follows from the monotonicity of the cdf (69) that the pdf is nonnegative

$$f_X(x) \geq 0, \tag{76}$$

since otherwise we would be able to find two points $x_1 < x_2$ for which $F_X(x_2) < F_X(x_1)$.

Remark 5.16 (The pdf is not a probability measure)**.** *The pdf is a* ***density*** *which must be integrated to yield a probability. In particular, it is not necessarily smaller than one (for example, take $a = 0$ and $b = 1/2$ in Definition 5.17 below).*

Finally, just as in the case of discrete random variables, we often say that a random variable is **distributed** according to a certain pdf or cdf, or that we know its distribution. The reason is that the pmf, pdf or cdf suffice to characterize the underlying probability space, as we mentioned before.

5.2.2 Important random variables

A **uniform** random variable models an experiment in which every outcome within a continuous interval is equally likely.

Definition 5.17 (Uniform)**.** *The pdf of a uniform random variable with domain* $[a, b]$ *is given by*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases} \tag{77}$$

The **exponential** random variable is often used to model *waiting times*: the time it takes until a certain event occurs. Examples of such events include the decay of a radioactive particle, a telephone call or the mechanical failure of a device.

Definition 5.18 (Exponential)**.** *The pdf of an exponential random variable with parameter* λ *is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{78}$$

An important property of an exponential random variable is that it is **memoryless**. Intuitively this means that knowing the time you have waited up to now gives you *no* information about how much more time you will have to wait (a feeling you might have felt calling customer service).

Lemma 5.19 (Exponential random variables are memoryless)**.** *Let* T *be an exponential random variable, for any* $t > t_0$

$$\mathrm{P}(T \leq t - t_0) = \mathrm{P}(T \leq t | T > t_0), \tag{79}$$

i.e. the waiting time starting at any t_0 *is distributed exactly like the original waiting time.*

Proof. From (74)

$$\mathrm{P}(T \leq t - t_0) = \int_0^{t-t_0} \lambda e^{-\lambda x} \mathrm{d}x = 1 - e^{-\lambda(t-t_0)}. \tag{80}$$

By the definition of conditional probability

$$\begin{aligned} \mathrm{P}(T \leq t | T > t_0) &= \frac{\mathrm{P}(t_0 \leq T \leq t)}{\mathrm{P}(T > t_0)} && (81) \\ &= \frac{\int_{t_0}^{t} \lambda e^{-\lambda x} \mathrm{d}x}{\int_{t_0}^{\infty} \lambda e^{-\lambda x} \mathrm{d}x} && (82) \\ &= \frac{e^{-\lambda t_0} - e^{-\lambda t}}{e^{-\lambda t_0}} = 1 - e^{-\lambda(t-t_0)}. && (83) \end{aligned}$$

□

The **Gaussian** or **normal** random variable is arguably the most notorious in probability and statistics. It is often used to model variables with unknown distributions in the natural sciences. This is motivated by the fact that sums of independent random variables converge to Gaussian distributions under certain assumptions. This phenomenon is captured by the Central Limit Theorem, which we will discuss further on in the course.

Definition 5.20 (Gaussian)**.** *The pdf of a Gaussian or normal random variable with parameters μ and σ is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{84}$$

It is not immediately obvious that the pdf of the Gaussian integrates to one.

Lemma 5.21. *The pdf of a Gaussian random variable integrates to one.*

Proof. The result is a consequence of the following lemma, proved in Section C of the appendix.

Lemma 5.22.

$$\int_{-\infty}^{\infty} e^{-t^2}\, dt = \sqrt{\pi}. \tag{85}$$

To complete the proof we use the change of variables $t = (x-\mu)/\sqrt{2}\sigma$. □

5.3 Functions of random variables

Within a probabilistic model an uncertain quantity Y may be well modeled as a function of another quantity X with known distribution: $Y = g(X)$. If X is discrete, by definition of the pmf,

$$\begin{aligned} p_Y(y) &= \mathrm{P}(Y = y) && (86)\\ &= \mathrm{P}(g(X) = y) && (87)\\ &= \sum_{\{x \mid g(x)=y\}} p_X(x). && (88) \end{aligned}$$

If X is continuous, then

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) && (89)\\ &= \mathrm{P}(g(X) \leq y) && (90)\\ &= \int_{\{x \mid g(x)\leq y\}} f_X(x)\,\mathrm{d}x, && (91) \end{aligned}$$

where the last equality only holds if the cdf of X is differentiable.

Example 5.23 (Voltage and power)**.** Your friend who is in electrical engineering has a good model for the voltage V across a resistor in a certain device: he has even been able to model its pdf f_V. He asks you what this implies about the behavior of the power X dissipated in the resistor (recall that $X = V^2/r$ where r is the resistance).

You quickly explain that for $x \geq 0$

$$F_X(x) = \int_{\{v \mid v^2/r \leq x\}} f_V(v)\, \mathrm{d}v \tag{92}$$

$$= \int_{v=-\sqrt{rx}}^{\sqrt{rx}} f_V(v)\, \mathrm{d}v \tag{93}$$

and

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d\ x}} = \frac{1}{2}\sqrt{\frac{r}{x}} \left(f_V\left(-\sqrt{rx}\right) + f_V\left(\sqrt{rx}\right) \right). \tag{94}$$

A Proof of Lemma 5.10

For any fixed constants c_1 and c_2

$$\lim_{n\to\infty} \frac{n-c_1}{n-c_2} = 1, \tag{95}$$

so that

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,(n-\lambda)^k} = \frac{n}{n-\lambda} \cdot \frac{n-1}{n-\lambda} \cdots \frac{n-k+1}{n-\lambda} = 1. \tag{96}$$

The result follows from the following basic calculus identity:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{97}$$

B Proof of Lemma 5.13

To establish (67)

$$\begin{aligned}
\lim_{x\to-\infty} F_X(x) &= 1 - \lim_{x\to-\infty} \mathrm{P}(X > x) && (98)\\
&= 1 - \mathrm{P}(X > 0) + \lim_{n\to\infty} \sum_{i=0}^{n} \mathrm{P}(-i \geq X > -(i+1)) && (99)\\
&= 1 - \mathrm{P}\left(\lim_{n\to\infty} \{X > 0\} \cup \cup_{i=0}^{n} \{-i \geq X > -(i+1)\}\right) \quad \text{by (2) in Definition 2.3} && (100)\\
&= 1 - \mathrm{P}(\Omega) = 0. && (101)
\end{aligned}$$

The proof of (68) follows from this result. Let $Y = -X$, then

$$\begin{aligned}
\lim_{x\to\infty} F_X(x) &= \lim_{x\to\infty} \mathrm{P}(X \leq x) && (102)\\
&= 1 - \lim_{x\to\infty} \mathrm{P}(X > x) && (103)\\
&= 1 - \lim_{x\to-\infty} \mathrm{P}(-X < x) && (104)\\
&= 1 - \lim_{x\to-\infty} F_Y(x) = 1 \quad \text{by (68).} && (105)
\end{aligned}$$

Finally, (69) holds by (8) because $\{X \leq a\} \subseteq \{X \leq b\}$.

C Proof of Lemma 5.22

Let us define

$$I = \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x. \tag{106}$$

Now taking the square and changing to polar coordinates,

$$\begin{aligned}
I^2 &= \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x \int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y && (107)\\
&= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{-\left(x^2+y^2\right)} \mathrm{d}x\mathrm{d}y && (108)\\
&= \int_{\theta=0}^{2\pi} \int_{r=-\infty}^{\infty} re^{-\left(r^2\right)} \mathrm{d}\theta\mathrm{d}r && (109)\\
&= \pi e^{-\left(r^2\right)}]_0^{\infty} = \pi. && (110)
\end{aligned}$$