Probability Theory Basics

1 Basic Definitions and Properties

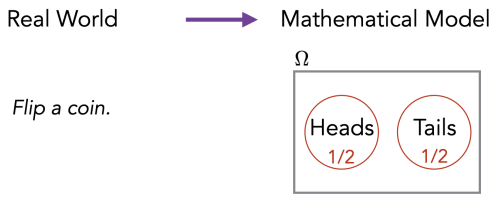
Definition (Finite probability space, sample space, probability distribution)**.** A *finite probability space* is a tuple $(\Omega, \mathbf{Pr})$, where

- Ω is a non-empty finite set called the *sample space*;
- $\mathbf{Pr} : \Omega \to [0, 1]$ is a function, called the *probability distribution*, with the property that $\sum_{\ell \in \Omega} \mathbf{Pr}[\ell] = 1$.

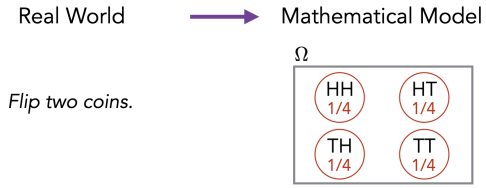
The elements of Ω are called *outcomes* or *samples*. If $\mathbf{Pr}[\ell] = p$, then we say that *the probability of outcome ℓ is p*.

Remark (Why do probabilities sum to 1?)**.** The probabilities sum to 1 by convention. We could have defined it so that they sum to 100, and think of probabilities in terms of percentages. Or we could have defined it so they sum to some other value. What is important is that a consistent choice is made. Note that the choice of 1 allows us to think of probabilities as fractions, and it is arguably the most natural choice.

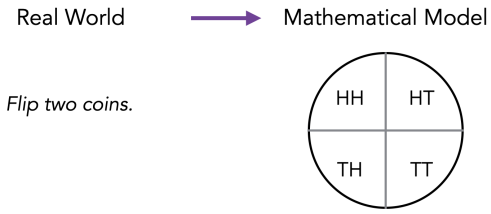
Note (Modeling randomness)**.** The abstract definition above of a finite probability space helps us to mathematically model and reason about situations involving randomness and uncertainties (these situations are often called "random experiments" or just "experiments"). For example, consider the experiment of flipping a single coin. We model this as follows. We let $\Omega = \{\text{Heads}, \text{Tails}\}$ and we define function $\mathbf{Pr}$ such that $\mathbf{Pr}[\text{Heads}] = 1/2$ and $\mathbf{Pr}[\text{Tails}] = 1/2$. This corresponds to our intuitive understanding that the probability of seeing the outcome Heads is $1/2$ and the probability of seeing the outcome Tails is also $1/2$.



If we flip two coins, the sample space would look as follows (H represents "Heads" and T represents "Tails").



One can visualize the probability space as a circle or pie with area 1. Each outcome gets a slice of the pie proportional to its probability.



Note (Restriction to finite sample spaces)**.** In this course, we'll usually restrict ourselves to finite sample spaces. In cases where we need a countably infinite Ω, the above definition will generalize naturally.

Exercise (Probability space modeling)**.** How would you model a roll of a single 6-sided die using Definition (Finite probability space, sample space, probability distribution)? How about a roll of two dice? How about a roll of a die and a coin toss together?

Solution. For the case of a single 6-sided die, we want the model to match our intuitive understanding and real-world experience that the probability of observing each of the possible die rolls $1, 2, \ldots, 6$ is equal. We formalize this by defining the *sample space* Ω and the *probability distribution* $\mathbf{Pr}$ as follows:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \qquad \mathbf{Pr}[\ell] = \frac{1}{6} \text{ for all } \ell \in \Omega.$$

Similarly, to model a roll of two dice, we can let each outcome be an ordered pair representing the roll of each of the two dice:

$$\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}, \qquad \mathbf{Pr}[\ell] = \left(\frac{1}{6}\right)^2 = \frac{1}{36} \text{ for all } \ell \in \Omega.$$

Lastly, to model a roll of a die and a coin toss together, we can let each outcome be an ordered pair where the first element represents the result of the die roll, and the second element represents the result of the coin toss:

$$\Omega = \{1, 2, 3, 4, 5, 6\} \times \{\text{Heads}, \text{Tails}\}, \qquad \mathbf{Pr}[\ell] = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12} \text{ for all } \ell \in \Omega.$$

■

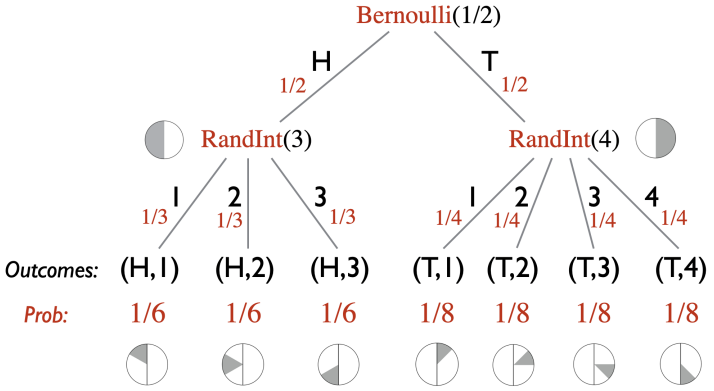
Note (Modeling through randomized code)**.** Sometimes, the modeling of a real-world random experiment as a probability space can be non-trivial or tricky. It helps a lot to have a step in between where you first go from the real-world experiment to computer code/algorithm (that makes calls to random number generators), and then you define your probability space based on the computer code. In this course, we allow our programs to have access to the functions Bernoulli(p) and RandInt(n). The function Bernoulli(p) takes a number $0 \leq p \leq 1$ as input and returns 1 with probability p and 0 with probability $1-p$. The function RandInt(n) takes a positive integer n as input and returns a random integer from 1 to n (i.e., every number from 1 to n has probability $1/n$). Here is a very simple example of going from a real-world experiment to computer code. The experiment is as follows. You flip a fair coin. If it's heads, you roll a 3-sided die. If it is tails, you roll a 4-sided die. This experiment can be represented as:

```
flip = Bernoulli(1/2)
if flip == 0:
    die = RandInt(3)
else:
    die = RandInt(4)
```

If we were to ask "What is the probability that you roll a 3 or higher?", this would correspond to asking what is the probability that after the above code is executed, the variable named `die` stores a value that is 3 or higher.

One advantage of modeling with randomized code is that if there is any ambiguity in the description of the random (real-world) experiment, then it would/should be resolved in this step of creating the randomized code.

The second advantage is that it allows you to easily imagine a *probability tree* associated with the randomized code. The probability tree gives you clear picture on what the sample space is and how to compute the probabilities of the outcomes. The probability tree corresponding to the above code is as follows.



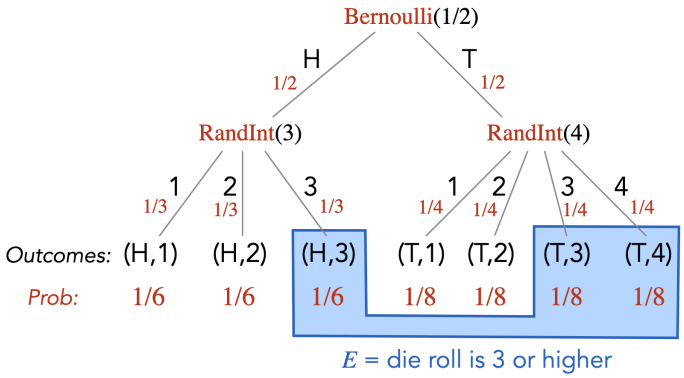
This simple example may not illustrate the usefulness of having a computer code representation of the random experiment, but one can appreciate its value with more sophisticated examples and we do encourage you to think of random experiments as computer code:

real-world experiment $\longrightarrow$ computer code / probability tree $\longrightarrow$ probability space $(\Omega, \mathbf{Pr})$.

Definition (Uniform distribution)**.** If a probability distribution $\mathbf{Pr} : \Omega \to [0,1]$ is such that $\mathbf{Pr}[\ell] = 1/|\Omega|$ for all $\ell \in \Omega$, then we call it a *uniform distribution*.

Definition (Event)**.** Let $(\Omega, \mathbf{Pr})$ be a probability space. Any subset of outcomes $E \subseteq \Omega$ is called an *event*. We extend the $\mathbf{Pr}[\cdot]$ notation and write $\mathbf{Pr}[E]$ to denote $\sum_{\ell \in E} \mathbf{Pr}[\ell]$. Using this notation, $\mathbf{Pr}[\varnothing] = 0$ and $\mathbf{Pr}[\Omega] = 1$. We use the notation $\overline{E}$ to denote the event $\Omega \backslash E$.

Example. Continuing the example given in Note (Modeling through randomized code), we can define the event $E = \{(H,3),(T,3),(T,4)\}$. In other words, E can be described as "die roll is 3 or higher". The probability of E, $\mathbf{Pr}[E]$, is equal to $1/6+1/8+1/8 = 5/12$.



Exercise (Practice with events).

1. Suppose we roll two 6-sided dice. How many events are there? Write down the event corresponding to "we roll a double" and determine its probability.
2. Suppose we roll a 3-sided die and see the number d. We then roll a d-sided die. How many different events are there? Write down the event corresponding to "the second roll is a 2" and determine its probability.

Solution. Part (1): We use the model for rolling two 6-sided dice as in Exercise (Probability space modeling). Since an event is any subset of outcomes $E \subseteq \Omega$, the number of events is the number of such subsets, which is $|\wp(\Omega)| = 2^{|\Omega|} = 2^{36}$ (here $\wp(\Omega)$ denotes the power set of Ω).

The event corresponding to "we roll a double" can be expressed as

$$E = \{(\ell,\ell) : \ell \in \{1,2,3,4,5,6\}\}$$

which has probability

$$\mathbf{Pr}[E] = \sum_{\ell \in E} \mathbf{Pr}[\ell] = \frac{1}{36} \cdot |E| = \frac{6}{36} = \frac{1}{6}.$$

Part(2): We model the two dice rolls as follows:

$$\Omega = \{(a,b) \in \{1,2,3\}^2 : a \geq b\}, \qquad \mathbf{Pr}[(a,b)] = \frac{1}{3} \cdot \frac{1}{a} = \frac{1}{3a}.$$

The restriction that $a \geq b$ is imposed because the second die depends on the first roll and the result of the second roll cannot be larger than that of the first. Note that we could also have used a model where $\Omega = \{1,2,3\}^2$ and assigned a probability of 0 to the outcomes where $a < b$, but considering outcomes that never occur is typically not very useful.

In the model we originally defined, the number of events is $|\wp(\Omega)| = 2^{|\Omega|} = 2^6 = 64$. Note that the number of events depends on the size of the sample space, so this number can vary depending on the model.

The event corresponding to "the second roll is a 2" is given by

$$E = \{(a,b) \in \Omega : b = 2\} = \{(2,2),(3,2)\}$$

which has probability

$$\mathbf{Pr}[E] = \sum_{\ell \in E} \mathbf{Pr}[\ell] = \mathbf{Pr}[(2,2)] + \mathbf{Pr}[(3,2)] = \frac{1}{3 \cdot 2} + \frac{1}{3 \cdot 3} = \frac{5}{18}.$$

■

Exercise (Basic facts about probability)**.** Let A and B be two events. Prove the following.

- If $A \subseteq B$, then $\mathbf{Pr}[A] \leq \mathbf{Pr}[B]$.
- $\mathbf{Pr}[\overline{A}] = 1 - \mathbf{Pr}[A]$.
- $\mathbf{Pr}[A \cup B] = \mathbf{Pr}[A] + \mathbf{Pr}[B] - \mathbf{Pr}[A \cap B]$.

Solution. Part (1): Suppose $A \subseteq B$. Recall that $\mathbf{Pr}$ is a non-negative function (i.e. $\mathbf{Pr}[\ell] \geq 0$ for all $\ell \in \Omega$). Hence,

$$\begin{aligned}
\mathbf{Pr}[A] &= \sum_{\ell \in A} \mathbf{Pr}[\ell] && \text{by the definition of event} \\
&\leq \sum_{\ell \in A} \mathbf{Pr}[\ell] + \sum_{\ell \in B \setminus A} \mathbf{Pr}[\ell] && \text{by non-negativity of } \mathbf{Pr} \\
&= \sum_{\ell \in B} \mathbf{Pr}[\ell] \\
&= \mathbf{Pr}[B].
\end{aligned}$$

Part (2): Recall that $\overline{A}$ denotes $\Omega \setminus A$, and that $\sum_{\ell \in \Omega} \mathbf{Pr}[\ell] = 1$. Hence,

$$\begin{aligned}
\mathbf{Pr}[\overline{A}] &= \sum_{\ell \in \overline{A}} \mathbf{Pr}[\ell] && \text{by definition of event} \\
&= \sum_{\ell \in \Omega \setminus A} \mathbf{Pr}[\ell] \\
&= \sum_{\ell \in \Omega} \mathbf{Pr}[\ell] - \sum_{\ell \in A} \mathbf{Pr}[\ell] \\
&= 1 - \mathbf{Pr}[A] && \text{by definition of prob. distr.}
\end{aligned}$$

Part (3) By partitioning $A \cup B$ into $A \setminus B$, $B \setminus A$, and $A \cap B$, we see that

$$\begin{aligned}
\mathbf{Pr}[A \cup B] &= \sum_{\ell \in A \cup B} \mathbf{Pr}[\ell] \quad \text{by definition of event} \\
&= \sum_{\ell \in A \setminus B} \mathbf{Pr}[\ell] + \sum_{\ell \in B \setminus A} \mathbf{Pr}[\ell] + \sum_{\ell \in A \cap B} \mathbf{Pr}[\ell] \\
&= \left(\sum_{\ell \in A} \mathbf{Pr}[\ell] - \sum_{\ell \in A \cap B} \mathbf{Pr}[\ell] \right) + \left(\sum_{\ell \in B} \mathbf{Pr}[\ell] - \sum_{\ell \in A \cap B} \mathbf{Pr}[\ell] \right) + \sum_{\ell \in A \cap B} \mathbf{Pr}[\ell] \\
&= \sum_{\ell \in A} \mathbf{Pr}[\ell] + \sum_{\ell \in B} \mathbf{Pr}[\ell] - \sum_{\ell \in A \cap B} \mathbf{Pr}[\ell] \\
&= \mathbf{Pr}[A] + \mathbf{Pr}[B] - \mathbf{Pr}[A \cap B] \quad \text{by definition of event.}
\end{aligned}$$

■

Definition (Disjoint events)**.** We say that two events A and B are *disjoint events* if $A \cap B = \varnothing$.

Exercise (Union bound)**.** Let $A_1, A_2, \ldots, A_n$ be events. Then

$$\mathbf{Pr}[A_1 \cup A_2 \cup \cdots \cup A_n] \leq \mathbf{Pr}[A_1] + \mathbf{Pr}[A_2] + \cdots + \mathbf{Pr}[A_n].$$

We get equality if and only if the A_i's are pairwise disjoint.

Solution. By the last part of Exercise (Basic facts about probability), we can conclude that

$$\mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B].$$

We also notice that equality holds if and only if $\mathbf{Pr}[A \cap B] = 0$, which happens if and only if A and B are disjoint. We will extend this by induction on n.

The expression is identical on both sides for the case where $n = 1$, and the case where $n = 2$ is exactly as above. These serve as the base cases for our induction.

For the inductive case, assume the proposition holds true for $n = k$. We seek to show that it too holds true for $n = k + 1$. Given events $A_1, A_2, \ldots, A_k, A_{k+1}$, let

$$A = A_1 \cup A_2 \cup \cdots \cup A_k, \qquad B = A_{k+1}.$$

Then

$$\begin{aligned}
\mathbf{Pr}[A_1 \cup A_2 \cup \cdots \cup A_k \cup A_{k+1}] &= \mathbf{Pr}[A \cup B] \\
&\leq \mathbf{Pr}[A] + \mathbf{Pr}[B] \quad \text{by the above result} \\
&= \mathbf{Pr}[A_1 \cup A_2 \cup \cdots \cup A_k] + \mathbf{Pr}[A_{k+1}] \\
&\leq (\mathbf{Pr}[A_1] + \cdots + \mathbf{Pr}[A_k]) + \mathbf{Pr}[A_{k+1}] \quad \text{by IH}
\end{aligned}$$

where equality holds if and only if A and B are disjoint and $A_1, \ldots, A_k$ are pairwise disjoint. In other words, equality holds if and only if

$$\bigcup_{t=1}^{k} (A_t \cap A_{k+1}) = \varnothing \qquad \text{and} \qquad A_i \cap A_j = \varnothing \text{ for all } 1 \leq i < j \leq k,$$

which in turn holds if and only if

$$A_i \cap A_j = \varnothing \text{ for all } 1 \leq i < j \leq k + 1.$$

(i.e. $A_1, \ldots, A_{k+1}$ are pairwise disjoint). This completes the proof. ■

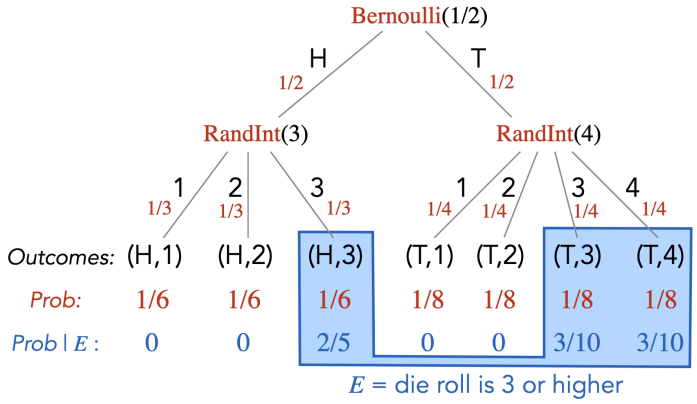
Definition (Conditional probability)**.** Let E be an event with $\mathbf{Pr}[E] \neq 0$. The *conditional probability* of outcome $\ell \in \Omega$ given E, denoted $\mathbf{Pr}[\ell \mid E]$, is defined as

$$\mathbf{Pr}[\ell \mid E] = \begin{cases} 0 & \text{if } \ell \notin E \\ \frac{\mathbf{Pr}[\ell]}{\mathbf{Pr}[E]} & \text{if } \ell \in E \end{cases}$$

For an event A, the *conditional probability of A given E*, denoted $\mathbf{Pr}[A \mid E]$, is defined as

$$\mathbf{Pr}[A \mid E] = \frac{\mathbf{Pr}[A \cap E]}{\mathbf{Pr}[E]}.$$

Example. Once again, continuing the example given in Note (Modeling through randomized code), let E be the event that the die roll is 3 or higher. Then for each outcome of the sample space, we can calculate its probability, given the event E.



For example, $\mathbf{Pr}[(H,1) \mid E] = 0$ and $\mathbf{Pr}[(H,3) \mid E] = 2/5$. We can also calculate the conditional probabilities of events. Let A be the event that the coin toss resulted in Tails. Then $\mathbf{Pr}[A \mid E] = 3/10 + 3/10 = 3/5$.

Note (Intuitive understanding of conditional probability)**.** Although it may not be immediately obvious, the above definition of conditional probability does correspond to our intuitive understanding of what conditional probability should represent. If we are told that event E has already happened, then we know that the probability of any outcome outside of E should be 0. Therefore, we can view the conditioning on event E as a transformation of our probability space where we revise the probabilities (i.e., we revise the probability function $\mathbf{Pr}[\cdot]$). In particular, the original probability space $(\Omega, \mathbf{Pr})$ gets transformed to $(\Omega, \mathbf{Pr}_E)$, where $\mathbf{Pr}_E$ is such that for any $\ell \notin E$, we have $\mathbf{Pr}_E[\ell] = 0$, and for any $\ell \in E$, we have $\mathbf{Pr}_E[\ell] = \mathbf{Pr}[\ell]/\mathbf{Pr}[E]$. The $1/\mathbf{Pr}[E]$ factor here is a necessary *normalization* factor that ensures the probabilities of all the outcomes sum to 1 (which is required by Definition (Finite probability space, sample space, probability distribution)). Indeed,

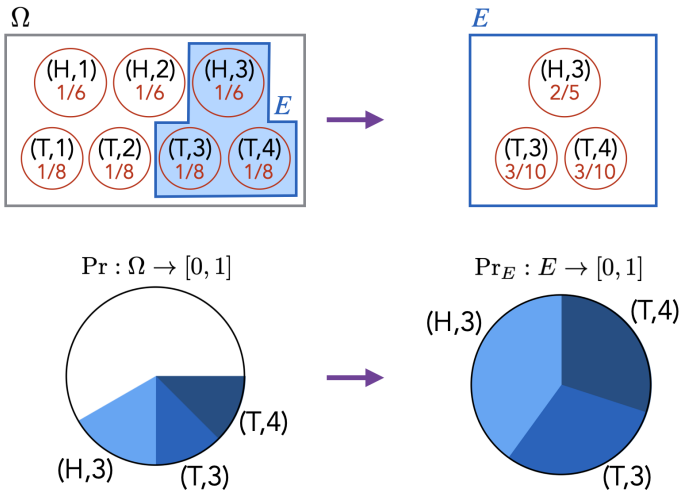
$$\begin{aligned}\textstyle\sum_{\ell\in\Omega} \mathbf{Pr}_E[\ell] &= \textstyle\sum_{\ell\notin E} \mathbf{Pr}_E[\ell] + \sum_{\ell\in E} \mathbf{Pr}_E[\ell] \\ &= 0 + \textstyle\sum_{\ell\in E} \mathbf{Pr}[\ell]/\mathbf{Pr}[E] \\ &= \frac{1}{\mathbf{Pr}[E]} \textstyle\sum_{\ell\in E} \mathbf{Pr}[\ell] \\ &= 1.\end{aligned}$$

If we are interested in the event "A given E" (denoted by $A \mid E$) in the probability space $(\Omega, \mathbf{Pr})$, then we are interested in the event A in the probability space $(\Omega, \mathbf{Pr}_E)$. That is, $\mathbf{Pr}[A \mid E] = \mathbf{Pr}_E[A]$. Therefore,

$$\mathbf{Pr}[A \mid E] = \mathbf{Pr}_E[A] = \mathbf{Pr}_E[A \cap E] = \frac{\mathbf{Pr}[A \cap E]}{\mathbf{Pr}[E]},$$

where the last equality holds by the definition of $\mathbf{Pr}_E[\cdot]$. We have thus recovered the equality in Definition (Conditional probability).

Conditioning on event E can also be viewed as redefining the sample space Ω to be E, and then renormalizing the probabilities so that $\mathbf{Pr}[\Omega] = \mathbf{Pr}[E] = 1$.



Exercise (Conditional probability practice)**.** Suppose we roll a 3-sided die and see the number d. We then roll a d-sided die. We are interested in the probability that the first roll was a 1 given that the second roll was a 1. First express this probability using the notation of conditional probability and then determine what the probability is.

Solution. We use the model defined in Exercise (Practice with events). The event that the first roll is a 1 is

$$E_1 = \{(1,1)\}$$

and the event that the second roll is a 1 is

$$E_2 = \{(1,1),(2,1),(3,1)\}$$

with corresponding probabilities

$$\mathbf{Pr}[E_1] = \frac{1}{3}, \qquad \mathbf{Pr}[E_2] = \frac{1}{3\cdot 1} + \frac{1}{3\cdot 2} + \frac{1}{3\cdot 3} = \frac{11}{18}.$$

Then the conditional probability we are interested in is

$$\mathbf{Pr}[E_1 \mid E_2] = \frac{\mathbf{Pr}[E_1 \cap E_2]}{\mathbf{Pr}[E_2]} = \frac{\mathbf{Pr}[E_1]}{\mathbf{Pr}[E_2]} = \frac{1/3}{11/18} = \frac{6}{11}.$$

■

Proposition (Chain rule). *Let $n \geq 2$ and let $A_1, A_2, \ldots, A_n$ be events. Then*

$$\mathbf{Pr}[A_1 \cap \cdots \cap A_n] = \\ \mathbf{Pr}[A_1] \cdot \mathbf{Pr}[A_2 \mid A_1] \cdot \mathbf{Pr}[A_3 \mid A_1 \cap A_2] \cdots \mathbf{Pr}[A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}].$$

Proof. We prove the proposition by induction on n. The base case with two events follows directly from the definition of conditional probability. Let $A = A_n$ and $B = A_1 \cap \ldots \cap A_{n-1}$. Then

$$\begin{aligned}\mathbf{Pr}[A_1 \cap \cdots \cap A_n] &= \mathbf{Pr}[A \cap B] \\ &= \mathbf{Pr}[B] \cdot \mathbf{Pr}[A \mid B] \\ &= \mathbf{Pr}[A_1 \cap \cdots \cap A_{n-1}] \cdot \mathbf{Pr}[A_n \mid A_1 \cap \cdots \cap A_{n-1}],\end{aligned}$$

where we used the definition of conditional probability for the second equality. Applying the induction hypothesis to $\mathbf{Pr}[A_1 \cap \cdots \cap A_{n-1}]$ gives the desired result. □

Exercise (Practice with chain rule). Suppose there are 100 students in 15-251 and 5 of the students are Andrew Yang supporters. We pick 3 students from class uniformly at random. Calculate the probability that none of them are Andrew Yang supporters using Proposition (Chain rule).

Solution. For $i = 1, 2, 3$, let A_i be the event that the i'th student we pick is not a Andrew Yang supporter. Then using the chain rule, the probability that none of them are Andrew Yang supporters is

$$\mathbf{Pr}[A_1 \cap A_2 \cap A_3] = \mathbf{Pr}[A_1] \cdot \mathbf{Pr}[A_2 \mid A_1] \cdot \mathbf{Pr}[A_3 \mid A_1 \cap A_2] = \frac{95}{100} \cdot \frac{94}{99} \cdot \frac{93}{98}.$$

■

Definition (Independent events).

- Let A and B be two events. We say that A and B are *independent events* if $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A] \cdot \mathbf{Pr}[B]$. Note that if $\mathbf{Pr}[B] \neq 0$, then this is equivalent to $\mathbf{Pr}[A \mid B] = \mathbf{Pr}[A]$. If $\mathbf{Pr}[A] \neq 0$, it is also equivalent to $\mathbf{Pr}[B \mid A] = \mathbf{Pr}[B]$.
- Let $A_1, A_2, \ldots, A_n$ be events. We say that $A_1, \ldots, A_n$ are *independent* if for any subset $S \subseteq \{1, 2, \ldots, n\}$,

$$\mathbf{Pr}\left[\bigcap_{i \in S} A_i\right] = \prod_{i \in S} \mathbf{Pr}[A_i].$$

Note (Defining independence through computer code)**.** Above we have given the definition of *independent events* as presented in 100% of the textbooks on probability theory. Yet, there is something deeply unsatisfying about this definition. In many situations people want to compute a probability of the form $\mathbf{Pr}[A \cap B]$, and if possible (if they are independent), would like to use the equality $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\,\mathbf{Pr}[B]$ to simplify the calculation. In order to do this, they will informally argue that events A and B are independent in the intuitive sense of the word. For example, they argue that if B happens, then this doesn't affect the probability of A happening (this argument is not done by calculation, but by informal argument). Then using this, they justify using the equality $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\,\mathbf{Pr}[B]$ in their calculations. So really, secretly, people are not using Definition (Independent events) but some other non-formal intuitive definition of independence, and then concluding what the formal definition says, which is $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\,\mathbf{Pr}[B]$.

To be a bit more explicit, recall that the approach to answering probability related questions is to go from a real-world experiment we want to analyze to a formal probability space model:

$$\text{real-world experiment} \longrightarrow \text{probability space } (\Omega, \mathbf{Pr}).$$

People often argue the independence of events A and B on the left-hand-side in order to use $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\,\mathbf{Pr}[B]$ on the right-hand-side. The left-hand-side, however, is not really a formal setting and may have ambiguities. And why does our intuitive notion of independence allow us to conclude $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\,\mathbf{Pr}[B]$? In these situations, it helps to add the "computer code" step in between:

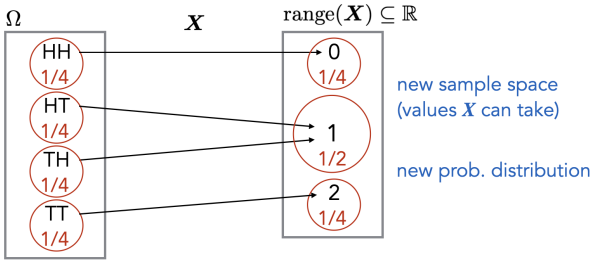
$$\text{real-world experiment} \longrightarrow \text{computer code} \longrightarrow \text{probability space } (\Omega, \mathbf{Pr}).$$

Computer code has no ambiguities and we can give a formal definition of independence using it. Suppose you have a randomized code modeling the real-world experiment, and suppose that you can divide the code into two separate parts. Suppose A is an event that depends only on the first part of the code, and B is an event that depends only on the second part of the code. If you can prove that the two parts of the code cannot affect each other, then we say that A and B are independent. When A and B are independent in this sense, then one can verify that indeed the equality $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\,\mathbf{Pr}[B]$ holds.

2 Random Variables Basics

Definition (Random variable)**.** A *random variable* is a function $\mathbf{X} : \Omega \to \mathbb{R}$, where Ω is a sample space.

Note (Random variable intuition)**.** Note that a random variable is just a labeling of the elements in Ω with some real numbers. One can think of this as a transformation of the original sample space into one that contains only numbers. For example, suppose the original sample space corresponds to flipping two coins. Then we can define a random variable $\mathbf{X}$ which maps an outcome in the sample space to the number of tails in the outcome. Since we are only flipping two coins, the possible outputs of $\mathbf{X}$ are 0, 1, and 2.



This kind of transformation is often desirable. For example, the transformation allows us to take a *weighted average* of the elements in the new sample space, where the weights correspond to the probabilities of the elements (if the distribution is uniform, the weighted average is just the regular average). This is called the *expectation* of the random variable and is formally defined below in Definition (Expected value of a random variable). Without this transformation into real numbers, the concept of an "expected value" (i.e. averaging) would not be possible to define.

Remark (Range of a random variable)**.** Almost always, the random variables we consider in this course will have a range that is a subset of $\mathbb{N}$.

Definition (Common events through a random variable)**.** Let $\mathbf{X}$ be a random variable and $x \in \mathbb{R}$ be some real value. We use

$$\begin{aligned}
\mathbf{X} = x &\quad \text{to denote the event } \{\ell \in \Omega : \mathbf{X}(\ell) = x\},\\
\mathbf{X} \leq x &\quad \text{to denote the event } \{\ell \in \Omega : \mathbf{X}(\ell) \leq x\},\\
\mathbf{X} \geq x &\quad \text{to denote the event } \{\ell \in \Omega : \mathbf{X}(\ell) \geq x\},\\
\mathbf{X} < x &\quad \text{to denote the event } \{\ell \in \Omega : \mathbf{X}(\ell) < x\},\\
\mathbf{X} > x &\quad \text{to denote the event } \{\ell \in \Omega : \mathbf{X}(\ell) > x\}.
\end{aligned}$$

For example, $\mathbf{Pr}[\mathbf{X} = x]$ denotes $\mathbf{Pr}[\{\ell \in \Omega : \mathbf{X}(\ell) = x\}]$. More generally, for $S \subseteq \mathbb{R}$, we use

$$\mathbf{X} \in S \quad \text{to denote the event } \{\ell \in \Omega : \mathbf{X}(\ell) \in S\}.$$

Exercise (Practice with random variables)**.** Suppose we roll two 6-sided dice. Let $\mathbf{X}$ be the random variable that denotes the sum of the numbers we see. Explicitly write down the input-output pairs for the function $\mathbf{X}$. Calculate $\mathbf{Pr}[\mathbf{X} \geq 7]$.

Solution. We use the model for rolling two 6-sided dice as in Exercise (Probability space modeling). Then

$$\begin{array}{llllll}
\mathbf{X}(1,1) = 2, & \mathbf{X}(1,2) = 3, & \mathbf{X}(1,3) = 4, & \mathbf{X}(1,4) = 5, & \mathbf{X}(1,5) = 6, & \mathbf{X}(1,6) = 7,\\
\mathbf{X}(2,1) = 3, & \mathbf{X}(2,2) = 4, & \mathbf{X}(2,3) = 5, & \mathbf{X}(2,4) = 6, & \mathbf{X}(2,5) = 7, & \mathbf{X}(2,6) = 8,\\
\mathbf{X}(3,1) = 4, & \mathbf{X}(3,2) = 5, & \mathbf{X}(3,3) = 6, & \mathbf{X}(3,4) = 7, & \mathbf{X}(3,5) = 8, & \mathbf{X}(3,6) = 9,\\
\mathbf{X}(4,1) = 5, & \mathbf{X}(4,2) = 6, & \mathbf{X}(4,3) = 7, & \mathbf{X}(4,4) = 8, & \mathbf{X}(4,5) = 9, & \mathbf{X}(4,6) = 10,\\
\mathbf{X}(5,1) = 6, & \mathbf{X}(5,2) = 7, & \mathbf{X}(5,3) = 8, & \mathbf{X}(5,4) = 9, & \mathbf{X}(5,5) = 10, & \mathbf{X}(5,6) = 11,\\
\mathbf{X}(6,1) = 7, & \mathbf{X}(6,2) = 8, & \mathbf{X}(6,3) = 9, & \mathbf{X}(6,4) = 10, & \mathbf{X}(6,5) = 11, & \mathbf{X}(6,6) = 12.
\end{array}$$

Since the probability distribution is uniform over the outcomes,

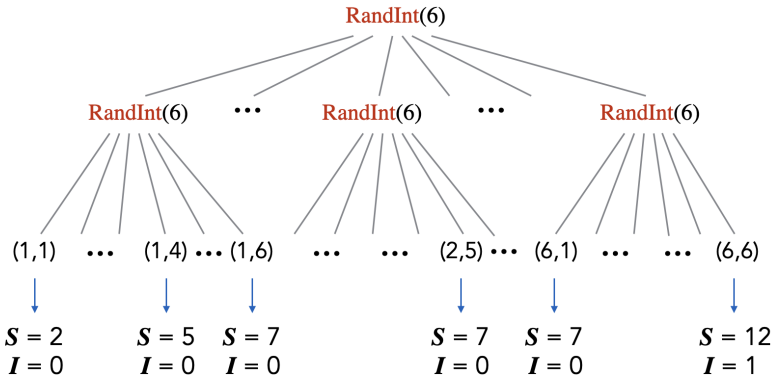
$$\mathbf{Pr}[\mathbf{X} \geq 7] = \frac{1}{36} \cdot |\{\ell \in \Omega : \mathbf{X}(\ell) \geq 7\}| = \frac{1}{36} \cdot 21 = \frac{7}{12}.$$

■

Note (Random variables and randomized code)**.** Using the randomized code and probability tree point of view, we can simply define a random variable as a numerical variable in some randomized code (more accurately, the variable's value at the end of the execution of the code). For example, consider the following randomized code.

```
S = RandInt(6) + RandInt(6)
if S == 12:
    I = 1
else:
    I = 0
```

Here we have two random variables corresponding to the variables in the code, **S** and **I**. From the probability tree picture, we can see how these two random variables can be viewed as functions from the sample space (the set of leaves) to $\mathbb{R}$ since they map each leaf/outcome to some numerical value.



Note (Forgetting the original sample space)**.** Given some probability space $(\Omega, \mathbf{Pr})$ and a random variable $\mathbf{X} : \Omega \to \mathbb{R}$, we often forget about the original sample space and consider the sample space to be the range of $\mathbf{X}$, $\text{range}(\mathbf{X}) = \{\mathbf{X}(\ell) : \ell \in \Omega\}$.

Definition (Probability mass function (PMF))**.** Let $\mathbf{X} : \Omega \to \mathbb{R}$ be a random variable. The *probability mass function* (*PMF*) of $\mathbf{X}$ is a function $p_{\mathbf{X}} : \mathbb{R} \to [0, 1]$ such that for any $x \in \mathbb{R}$, $p_{\mathbf{X}}(x) = \mathbf{Pr}[\mathbf{X} = x]$.

Note (Defining a random variable through PMF)**.** Related to the previous remark, we sometimes "define" a random variable by just specifying its probability mass function. In particular, we make no mention of the underlying sample space.

Definition (Expected value of a random variable)**.** Let $\mathbf{X}$ be a random variable. The *expected value* of $\mathbf{X}$, denoted $\mathbf{E}[\mathbf{X}]$, is defined as follows:

$$\mathbf{E}[\mathbf{X}] = \sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot \mathbf{X}(\ell).$$

Equivalently,

$$\mathbf{E}[\mathbf{X}] = \sum_{x \in \text{range}(\mathbf{X})} \mathbf{Pr}[\mathbf{X} = x] \cdot x,$$

where $\text{range}(\mathbf{X}) = \{\mathbf{X}(\ell) : \ell \in \Omega\}$.

Exercise (Equivalence of expected value definitions)**.** Prove that the above two expressions for $\mathbf{E}[\mathbf{X}]$ are equivalent.

Solution. We show a chain of equalities from the RHS to the LHS:

$$\begin{aligned}
\sum_{x \in \text{range}(\mathbf{X})} \mathbf{Pr}[\mathbf{X} = x] \cdot x &= \sum_{x \in \text{range}(\mathbf{X})} \mathbf{Pr}[\{\ell \in \Omega : \mathbf{X}(\ell) = x\}] \cdot x \\
&= \sum_{x \in \text{range}(\mathbf{X})} \sum_{\substack{\ell \in \Omega \\ \mathbf{X}(\ell) = x}} \mathbf{Pr}[\ell] \cdot x \\
&= \sum_{x \in \text{range}(\mathbf{X})} \sum_{\substack{\ell \in \Omega \\ \mathbf{X}(\ell) = x}} \mathbf{Pr}[\ell] \cdot \mathbf{X}(\ell) \\
&= \sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot \mathbf{X}(\ell).
\end{aligned}$$

■

Exercise (Practice with expected value)**.** Suppose we roll two 6-sided dice. Let $\mathbf{X}$ be the random variable that denotes the sum of the numbers we see. Calculate $\mathbf{E}[\mathbf{X}]$.

Solution. We refer to the input-output pairs of $\mathbf{X}$ (recall that random variables are functions) in Exercise (Practice with random variables). With a lot of tedious calculations, we can compute

$$\mathbf{E}[\mathbf{X}] = \sum_{x \in \text{range}(\mathbf{X})} \mathbf{Pr}[\mathbf{X} = x] \cdot x = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \cdots + \frac{1}{36} \cdot 12 = 7.$$

We will see a less tedious way of performing such calculations in the following exercise. ■

Proposition (Linearity of expectation)**.** *Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables, and let $c_1, c_2 \in \mathbb{R}$ be some constants. Then*

$$\mathbf{E}[c_1\mathbf{X} + c_2\mathbf{Y}] = c_1\,\mathbf{E}[\mathbf{X}] + c_2\,\mathbf{E}[\mathbf{Y}].$$

Proof. Define the random variable $\mathbf{Z}$ as $\mathbf{Z} = c_1\mathbf{X} + c_2\mathbf{Y}$. Then using the definition of expected value, we have

$$\begin{aligned}
\mathbf{E}[c_1\mathbf{X} + c_2\mathbf{Y}] &= \mathbf{E}[\mathbf{Z}] \\
&= \sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot \mathbf{Z}(\ell) \\
&= \sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot (c_1\mathbf{X}(\ell) + c_2\mathbf{Y}(\ell)) \\
&= \sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot c_1\mathbf{X}(\ell) + \mathbf{Pr}[\ell] \cdot c_2\mathbf{Y}(\ell) \\
&= \left(\sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot c_1\mathbf{X}(\ell)\right) + \left(\sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot c_2\mathbf{Y}(\ell)\right) \\
&= c_1\left(\sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot \mathbf{X}(\ell)\right) + c_2\left(\sum_{\ell \in \Omega} \mathbf{Pr}[\ell] \cdot \mathbf{Y}(\ell)\right) \\
&= c_1\,\mathbf{E}[\mathbf{X}] + c_2\,\mathbf{E}[\mathbf{Y}],
\end{aligned}$$

as desired. □

Corollary (Linearity of expectation 2)**.** *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be random variables, and $c_1, c_2, \ldots, c_n \in \mathbb{R}$ be some constants. Then*

$$\mathbf{E}[c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \cdots + c_n\mathbf{X}_n] = c_1\,\mathbf{E}[\mathbf{X}_1] + c_2\,\mathbf{E}[\mathbf{X}_2] + \cdots + c_n\,\mathbf{E}[\mathbf{X}_n].$$

In particular, when all the c_i's are 1, we get

$$\mathbf{E}[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n] = \mathbf{E}[\mathbf{X}_1] + \mathbf{E}[\mathbf{X}_2] + \cdots + \mathbf{E}[\mathbf{X}_n].$$

Exercise (Practice with linearity of expectation)**.** Suppose we roll three 10-sided dice. Let $\mathbf{X}$ be the sum of the three values we see. Calculate $\mathbf{E}[\mathbf{X}]$.

Solution. Let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ be the values of the rolls of each of the three dice. Note that $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are random variables and that $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$. Then since $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are

identically distributed, we can compute

$$\begin{aligned}\mathbf{E}[\mathbf{X}_1] = \mathbf{E}[\mathbf{X}_2] = \mathbf{E}[\mathbf{X}_3] &= \sum_{x \in \text{range}(\mathbf{X}_1)} \mathbf{Pr}[\mathbf{X}_1 = x] \cdot x \\ &= \sum_{x=1}^{10} \mathbf{Pr}[\mathbf{X}_1 = x] \cdot x \qquad \text{since } \text{range}(\mathbf{X}_1) = \{1, 2, \ldots, 10\} \\ &= \sum_{x=1}^{10} \frac{1}{10} \cdot x \\ &= \frac{11}{2}\end{aligned}$$

and by linearity of expectation,

$$\mathbf{E}[\mathbf{X}] = \mathbf{E}[\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3] = \mathbf{E}[\mathbf{X}_1] + \mathbf{E}[\mathbf{X}_2] + \mathbf{E}[\mathbf{X}_3] = 3 \cdot \frac{11}{2} = \frac{33}{2}.$$

■

Definition (Independent random variables)**.** Two random variables $\mathbf{X}$ and $\mathbf{Y}$ are *independent random variables* if for all $x, y \in \mathbb{R}$, the events $\mathbf{X} = x$ and $\mathbf{Y} = y$ are independent. The definition generalizes to more than two random variables analogous to Definition (Independent events).

Proposition (Expectation of product of independent random variables)**.** *If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent random variables, then*

$$\mathbf{E}[\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_n] = \mathbf{E}[\mathbf{X}_1] \cdot \mathbf{E}[\mathbf{X}_2] \cdots \mathbf{E}[\mathbf{X}_n].$$

Proof. We will first show the following sub-claim: if $\mathbf{X}$ and $\mathbf{Y}$ are independent random variables, then

$$\mathbf{E}[\mathbf{XY}] = \mathbf{E}[\mathbf{X}] \cdot \mathbf{E}[\mathbf{Y}].$$

Indeed, noting that the events $(\mathbf{X} = x) \cap (\mathbf{Y} = y)$, for $x \in \text{range}(\mathbf{X})$ and $y \in \text{range}(\mathbf{Y})$, partition Ω,

$$\begin{aligned}\mathbf{E}[\mathbf{XY}] &= \sum_{x \in \text{range}(\mathbf{X})} \sum_{y \in \text{range}(\mathbf{Y})} \mathbf{E}[\mathbf{XY} \mid (\mathbf{X} = x) \cap (\mathbf{Y} = y)] \cdot \mathbf{Pr}[(\mathbf{X} = x) \cap (\mathbf{Y} = y)] \\ &= \sum_{x \in \text{range}(\mathbf{X})} \sum_{y \in \text{range}(\mathbf{Y})} xy \cdot \mathbf{Pr}[(\mathbf{X} = x) \cap (\mathbf{Y} = y)] \\ &= \sum_{x \in \text{range}(\mathbf{X})} \sum_{y \in \text{range}(\mathbf{Y})} xy \cdot \mathbf{Pr}[\mathbf{X} = x] \cdot \mathbf{Pr}[\mathbf{Y} = y] \\ &= \left(\sum_{x \in \text{range}(\mathbf{X})} x \cdot \mathbf{Pr}[\mathbf{X} = x] \right) \cdot \left(\sum_{y \in \text{range}(\mathbf{Y})} y \cdot \mathbf{Pr}[\mathbf{Y} = y] \right) \\ &= \mathbf{E}[\mathbf{X}] \cdot \mathbf{E}[\mathbf{Y}].\end{aligned}$$

Now, we prove the original statement by induction on n. Both sides are identical for the case where $n = 1$, and the above claim is exactly the case where $n = 2$. These are our base cases.

For the inductive case, assume the proposition holds true for $n = k$. We seek to show that it also holds for $n = k + 1$. Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k+1}$ are independent random

variables. Let $\mathbf{X} = \prod_{j=1}^{k} \mathbf{X}_j$ and $\mathbf{Y} = \mathbf{X}_{k+1}$. We have

$$\begin{aligned}
\mathbf{E}\left[\prod_{j=1}^{k+1} \mathbf{X}_j\right] &= \mathbf{E}[\mathbf{XY}] \\
&= \mathbf{E}[\mathbf{X}] \cdot \mathbf{E}[\mathbf{Y}] \quad \text{by the above claim} \\
&= \mathbf{E}\left[\prod_{j=1}^{k} \mathbf{X}_j\right] \cdot \mathbf{E}[\mathbf{X}_{k+1}] \\
&= \left(\prod_{j=1}^{k} \mathbf{E}[\mathbf{X}_j]\right) \cdot \mathbf{E}[\mathbf{X}_{k+1}] \quad \text{by IH} \\
&= \prod_{j=1}^{k+1} \mathbf{E}[\mathbf{X}_j]
\end{aligned}$$

which completes the proof. □

Definition (Indicator random variable)**.** Let $E \subseteq \Omega$ be some event. The *indicator random variable* with respect to E is denoted by $\mathbf{I}_E$ and is defined as

$$\mathbf{I}_E(\ell) = \begin{cases} 1 & \text{if } \ell \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Proposition (Expectation of an indicator random variable)**.** *Let E be an event. Then* $\mathbf{E}[\mathbf{I}_E] = \mathbf{Pr}[E]$.

Proof. By the definition of expected value,

$$\begin{aligned}
\mathbf{E}[\mathbf{I}_E] &= \mathbf{Pr}[\mathbf{I}_E = 1] \cdot 1 + \mathbf{Pr}[\mathbf{I}_E = 0] \cdot 0 \\
&= \mathbf{Pr}[\mathbf{I}_E = 1] \\
&= \mathbf{Pr}[\{\ell \in \Omega : \mathbf{I}_E(\ell) = 1\}] \\
&= \mathbf{Pr}[\{\ell \in \Omega : \ell \in E\}] \\
&= \mathbf{Pr}[E].
\end{aligned}$$

□

Important (Combining linearity of expectation and indicators)**.** Suppose that you are interested in computing $\mathbf{E}[\mathbf{X}]$ for some random variable $\mathbf{X}$. If you can write $\mathbf{X}$ as a sum of indicator random variables, i.e., if $\mathbf{X} = \sum_j \mathbf{I}_{E_j}$ where $\mathbf{I}_{E_j}$ are indicator random variables, then by linearity of expectation,

$$\mathbf{E}[\mathbf{X}] = \mathbf{E}\left[\sum_j \mathbf{I}_{E_j}\right] = \sum_j \mathbf{E}[\mathbf{I}_{E_j}].$$

Furthermore, by Proposition (Expectation of an indicator random variable), we know $\mathbf{E}[\mathbf{I}_{E_j}] = \mathbf{Pr}[E_j]$. Therefore $\mathbf{E}[\mathbf{X}] = \sum_j \mathbf{Pr}[E_j]$. This often provides an extremely convenient way of computing $\mathbf{E}[\mathbf{X}]$. This combination of indicator random variables together with linearity expectation is one of the most useful tricks in probability theory!

Exercise (Practice with linearity of expectation and indicators)**.**

1. There are n balls and n bins. For each ball, you pick one of the bins uniformly at random and drop the ball in that bin. What is the expected number of balls in bin 1? What is the expected number of empty bins?

2. Suppose you randomly color the vertices of the complete graph on n vertices one of k colors. What is the expected number of paths of length c (where we assume $c \geq 3$) such that no two adjacent vertices on the path have the same color?

Solution. Part (1): Let $\mathbf{X}$ be the number of balls in bin 1. For $j = 1, 2, \ldots, n$, let E_j be the event that the j'th ball is dropped in bin 1. Observe that $\mathbf{X} = \sum_{j=1}^{n} \mathbf{I}_{E_j}$, and that $\mathbf{Pr}[E_j] = 1/n$ for all j, since the bin each ball is dropped into is picked uniformly at random. Then by linearity of expectation,

$$\mathbf{E}[\mathbf{X}] = \mathbf{E}\left[\sum_{j=1}^{n} \mathbf{I}_{E_j}\right] = \sum_{j=1}^{n} \mathbf{E}[\mathbf{I}_{E_j}] = \sum_{j=1}^{n} \mathbf{Pr}[E_j] = \sum_{j=1}^{n} \frac{1}{n} = 1.$$

Let $\mathbf{Y}$ be the number of empty bins. For $j = 1, 2, \ldots, n$, let F_j be the event that bin j is empty. Observe that $\mathbf{Y} = \sum_{j=1}^{n} \mathbf{I}_{F_j}$, and that $\mathbf{Pr}[F_j] = (1 - 1/n)^n$ for all j, since the probability that any one of the n balls is *not* dropped in a fixed bin is $1 - 1/n$, and each ball is dropped independently of the others. Then by linearity of expectation,

$$\mathbf{E}[\mathbf{Y}] = \mathbf{E}\left[\sum_{j=1}^{n} \mathbf{I}_{F_j}\right] = \sum_{j=1}^{n} \mathbf{E}[\mathbf{I}_{F_j}] = \sum_{j=1}^{n} \mathbf{Pr}[F_j] = \sum_{j=1}^{n} \left(1 - \frac{1}{n}\right)^n = n\left(1 - \frac{1}{n}\right)^n.$$

Part (2): Let $\mathbf{X}$ be the number of paths of length c such that no two adjacent vertices on the path have the same color. There are a total of

$$n(n-1)\cdots(n-c) = \frac{n!}{(n-c-1)!}$$

paths of length c. Let's call this value N and let's number the paths from 1 to N. For $j = 1, 2, \ldots, N$, let E_j be the event that no two adjacent vertices on the j'th path have the same color. Note that $\mathbf{X} = \sum_{j=1}^{N} \mathbf{I}_{E_j}$.

We first compute $\mathbf{Pr}[E_j]$ for some fixed j. Suppose path j is $v_0 v_1 \cdots v_c$. Then E_j occurs if and only if v_i is colored differently from v_{i-1} for $i = 1, 2, \ldots, c$. For each i, this happens with probability $1 - 1/k$, and they are independent of each other. Hence, we can conclude that $\mathbf{Pr}[E_j] = (1 - 1/k)^c$ for each j. Then by linearity of expectation,

$$\mathbf{E}[\mathbf{X}] = \mathbf{E}\left[\sum_{j=1}^{N} \mathbf{I}_{E_j}\right] = \sum_{j=1}^{N} \mathbf{E}[\mathbf{I}_{E_j}] = \sum_{j=1}^{N} \mathbf{Pr}[E_j] = \sum_{j=1}^{N} \left(1 - \frac{1}{k}\right)^c = \frac{n!}{(n-c-1)!}\left(1 - \frac{1}{k}\right)^c.$$

■

Theorem (Markov's inequality)**.** *Let $\mathbf{X}$ be a non-negative random variable. Then for any $c > 0$,*

$$\mathbf{Pr}[\mathbf{X} \geq c \cdot \mathbf{E}[\mathbf{X}]] \leq \frac{1}{c}.$$

Or equivalently,

$$\mathbf{Pr}[\mathbf{X} \geq c] \leq \frac{\mathbf{E}[\mathbf{X}]}{c}.$$

Remark (Markov's inequality intuition)**.** At a high level, Markov's inequality states that a non-negative random variable is rarely much bigger than its expectation.

For example, if the average score on an exam is 45%, then the fraction of the students who got at least 90% cannot be very large. And we can put an upper bound on that fraction by considering the scenario that would maximize the fraction of students getting at least 90%. To maximize that fraction, the contribution to the average score, of students who got below 90%, should be minimized. For this, we'll assume that anyone who got below a 90% actually got a 0% on the exam. Furthermore, anyone who gets strictly above

a 90% (e.g. anyone who gets a 100%) would reduce the fraction of students who get at least 90%. So we'll assume that anyone who got at least 90% actually got exactly 90%. In this scenario, if p is the fraction of students who got 90% on the exam, the average score on the exam would be $p \cdot 90$. This quantity should be equal to the actual average, 45. So p is equal to $1/2$. Meaning, if the average score in the exam is 45%, then at most half the class can get a score of at least 90%.

A more succinct way to say the above is that, if p is the fraction of students who got at least 90%, then the average must be at least $p \cdot 90$. So $50 \geq p \cdot 90$, and therefore $p \leq 1/2$.

Since we are not assuming anything about the random variable other than it is non-negative with non-zero expectation, the bound that Markov's inequality gives us is rather weak. There are much stronger bounds for more specific random variables.

Proof. Our goal is to find an upper bound on the probability that the random variable $\mathbf{X}$ is greater than or equal to $c \cdot \mathbf{E}[\mathbf{X}]$. Let p be the probability of this event. We want to show $p \leq 1/c$. We will do so as follows. First, we put a lower bound on $\mathbf{E}[\mathbf{X}]$. Recall that

$$\mathbf{E}[\mathbf{X}] = \sum_{x \in \text{range}(\mathbf{X})} \mathbf{Pr}[\mathbf{X} = x] \cdot x.$$

We divide this sum into two parts, based on whether $x \geq c \cdot \mathbf{E}[\mathbf{X}]$ or not:

$$\mathbf{E}[\mathbf{X}] = \left(\sum_{x < c \cdot \mathbf{E}[\mathbf{X}]} \mathbf{Pr}[\mathbf{X} = x] \cdot x \right) + \left(\sum_{x \geq c \cdot \mathbf{E}[\mathbf{X}]} \mathbf{Pr}[\mathbf{X} = x] \cdot x \right).$$

Using the non-negativity of $\mathbf{X}$, we know the first sum is at least 0. The second sum can be lower-bounded by $p \cdot c \cdot \mathbf{E}[\mathbf{X}]$ (because in the worst case, all the probability mass p is on the event $\mathbf{X} = x$ where $x = c \cdot \mathbf{E}[\mathbf{X}]$). Therefore, $\mathbf{E}[\mathbf{X}] \geq p \cdot c \cdot \mathbf{E}[\mathbf{X}]$. Rearranging, we get $p \leq 1/c$, as desired. □

Exercise (Practice with Markov's inequality)**.** During the Spring 2022 semester, the 15-251 TAs decide to strike because they are not happy with the lack of free food in grading sessions. Without the TA support, the performance of the students in the class drop dramatically. The class average on the first midterm exam is 15%. Using Markov's Inequality, give an upper bound on the fraction of the class that got an A (i.e., at least a 90%) in the exam.

Solution. Let $\mathbf{X}$ be the exam score of a student chosen uniformly at random. We will optimistically assume that $\mathbf{X}$ is non-negative. Then $\mathbf{E}[\mathbf{X}] = 0.15 \neq 0$ and the fraction of the class that got an A is $\mathbf{Pr}[\mathbf{X} \geq 0.9]$. By Markov's inequality,

$$\mathbf{Pr}[\mathbf{X} \geq 0.9] \leq \frac{\mathbf{E}[\mathbf{X}]}{0.9} = \frac{1}{6}$$

which is an upper bound on the fraction of the class that got an A. ■

3 Three Popular Random Variables

Definition (Bernoulli random variable)**.** Let $0 < p < 1$ be some parameter. If $\mathbf{X}$ is a random variable with probability mass function $p_{\mathbf{X}}(1) = p$ and $p_{\mathbf{X}}(0) = 1 - p$, then we say that $\mathbf{X}$ has a *Bernoulli distribution with parameter* p (we also say that $\mathbf{X}$ is a Bernoulli random variable). We write $\mathbf{X} \sim \text{Bernoulli}(p)$ to denote this. The parameter p is often called the *success* probability.

Note (What does a Bernoulli random variable represent?)**.** A Bernoulli random variable Bernoulli(p) captures a random experiment where we toss a p-biased coin where the probability of heads is p (and we assign this the numerical outcome of 1) and the probability of tails is $1 - p$ (and we assign this the numerical outcome of 0).

Note (Expectation of Bernoulli random variable)**.** Note that $\mathbf{E}[\mathbf{X}] = 0 \cdot p_{\mathbf{X}}(0) + 1 \cdot p_{\mathbf{X}}(1) = p_{\mathbf{X}}(1) = p$.

Definition (Binomial random variable)**.** Let $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$, where the $\mathbf{X}_i$'s are independent and for all i, $\mathbf{X}_i \sim$ Bernoulli(p). Then we say that $\mathbf{X}$ has a *binomial distribution with parameters n and p* (we also say that $\mathbf{X}$ is a binomial random variable). We write $\mathbf{X} \sim \text{Bin}(n, p)$ to denote this.

Note (What does a Binomial random variable represent?)**.** A Binomial random variable $\mathbf{X} \sim \text{Bin}(n, p)$ captures a random experiment where we toss a p-biased coin n times. We are interested in the probability of seeing k heads among those n coin tosses, where k ranges over $\{0, 1, 2, \ldots, n\} = \text{range}(\mathbf{X})$.

Note (Bernoulli is a special case of Binomial)**.** Note that we can view a Bernoulli random variable as a special kind of a binomial random variable where $n = 1$.

Exercise (Expectation of a Binomial random variable)**.** Let $\mathbf{X}$ be a random variable with $\mathbf{X} \sim \text{Bin}(n, p)$. Determine $\mathbf{E}[\mathbf{X}]$ (use linearity of expectation). Also determine $\mathbf{X}$'s probability mass function.

Solution. By definition, $\mathbf{X} = \sum_{j=1}^{n} \mathbf{X}_j$ where the $\mathbf{X}_i$'s are independent and for all i, $\mathbf{X}_i \sim$ Bernoulli(p). Since $\mathbf{E}[\mathbf{X}_i] = p$, by linearity of expectation:

$$\mathbf{E}[\mathbf{X}] = \mathbf{E}\left[\sum_{j=1}^{n} \mathbf{X}_j\right] = \sum_{j=1}^{n} \mathbf{E}\left[\mathbf{X}_j\right] = np.$$

We now determine the probability mass function of $\mathbf{X}$. Note that the values $\mathbf{X}$ can take on are $\{0, 1, 2, \ldots, n\}$. Let k be an arbitrary element of $\{0, 1, 2, \ldots, n\}$. Then,

$$\begin{aligned}
p_{\mathbf{X}}(k) &= \mathbf{Pr}[\mathbf{X} = k] \\
&= \mathbf{Pr}\left[\sum_{j=1}^{n} \mathbf{X}_j = k\right] \\
&= \sum_{\substack{J \subseteq \{1,\ldots,n\} \\ |J| = k}} \mathbf{Pr}\left[\bigcap_{j \in J} (\mathbf{X}_j = 1) \cap \bigcap_{j \notin J} (\mathbf{X}_j = 0)\right] \\
&= \sum_{\substack{J \subseteq \{1,\ldots,n\} \\ |J| = k}} \left(\prod_{j \in J} \mathbf{Pr}[\mathbf{X}_j = 1] \cdot \prod_{j \notin J} \mathbf{Pr}[\mathbf{X}_j = 0]\right) \\
&= \sum_{\substack{J \subseteq \{1,\ldots,n\} \\ |J| = k}} \left(p^k \cdot (1-p)^{n-k}\right) \\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

■

Exercise (Practice with Binomial random variable)**.** We toss a coin 5 times. What is the probability that we see at least 4 heads? What is the expected number of heads?

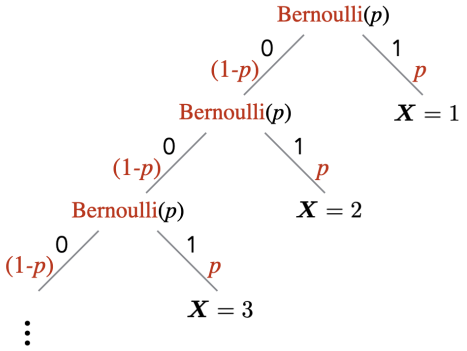
Solution. Let $\mathbf{X}$ be the number of heads among the 5 coin tosses. Then $\mathbf{X} \sim$ Binomial$(5, 1/2)$ and so the probability we see at least 4 heads is

$$\mathbf{Pr}[\mathbf{X} \geq 4] = \mathbf{Pr}[\mathbf{X} = 4] + \mathbf{Pr}[\mathbf{X} = 5] = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{3}{16}.$$

The expected number of heads, $\mathbf{E}[\mathbf{X}]$, is equal to $5 \cdot \frac{1}{2} = 2.5$. ■

Definition (Geometric random variable)**.** Let $\mathbf{X}$ be a random variable with probability mass function $p_{\mathbf{X}}$ such that for $n \in \{1, 2, \ldots\}$, $p_{\mathbf{X}}(n) = (1-p)^{n-1}p$. Then we say that $\mathbf{X}$ has a *geometric distribution with parameter* p (we also say that $\mathbf{X}$ is a geometric random variable). We write $\mathbf{X} \sim \text{Geometric}(p)$ to denote this.

Note (What does a Geometric random variable represent?)**.** A Geometric random variable Geometric(p) captures a random experiment where we successively toss a p-biased coin until we see heads for the first time, and we stop. We are interested in the probability of making n coin tosses in total before we stop, where n ranges over $\{1, 2, \ldots\}$.



Exercise (PMF of a geometric random variable)**.** Let $\mathbf{X}$ be a geometric random variable. Verify that $\sum_{n=1}^{\infty} p_{\mathbf{X}}(n) = 1$.

Solution. Recall that for $|r| < 1$, $\sum_{n=0}^{\infty} r^n = 1/(1-r)$. Then

$$\sum_{n=1}^{\infty} p_{\mathbf{X}}(n) = \sum_{n=1}^{\infty} (1-p)^{n-1}p = p \cdot \sum_{n=0}^{\infty} (1-p)^n = p \cdot \frac{1}{1-(1-p)} = 1.$$

■

Exercise (Practice with geometric random variable)**.** Suppose we repeatedly flip a coin until we see a heads for the first time. What is the probability that we will flip the coin more than 5 times?

Solution. Let $\mathbf{X}$ be the number of flips until we see a heads for the first time. Then $\mathbf{X} \sim \text{Geometric}(1/2)$. The question asks us to compute $\mathbf{Pr}[\mathbf{X} > 5]$. We know that the event $\mathbf{X} > 5$ is equivalent to getting tails in our first 5 flips, which happens with probability $1/2^5$. Therefore $\mathbf{Pr}[\mathbf{X} > 5] = 1/2^5$. ■

Exercise (Expectation of a geometric random variable)**.** Let $\mathbf{X}$ be a random variable with $\mathbf{X} \sim \text{Geometric}(p)$. Determine $\mathbf{E}[\mathbf{X}]$.

Solution. By the definition of expectation, we have $\mathbf{E}[\mathbf{X}] = \sum_{n=1}^{\infty} n \cdot (1-p)^{n-1}p$. Using

the fact that $\sum_{n=0}^{\infty}(1-p)^n = 1/p$, we get:

$$\begin{aligned}
\mathbf{E}[\mathbf{X}] &= \sum_{n=1}^{\infty} n \cdot (1-p)^{n-1} p \\
&= p\left(\sum_{n=1}^{\infty} n \cdot (1-p)^{n-1}\right) \\
&= p\left(\sum_{n=1}^{\infty}(1-p)^{n-1} + \sum_{n=2}^{\infty}(1-p)^{n-1} + \sum_{n=3}^{\infty}(1-p)^{n-1} + \cdots\right) \\
&= p\left(\frac{1}{p} + \frac{1-p}{p} + \frac{(1-p)^2}{p} + \cdots\right) \\
&= 1 + (1-p) + (1-p)^2 + \cdots \\
&= \frac{1}{p}.
\end{aligned}$$

■

Important (Some general tips)**.** Here are some general tips on probability calculations (this is not meant to be an exhaustive list).

- If you are trying to upper bound $\mathbf{Pr}[A]$, you can try to find B with $A \subseteq B$, and then bound $\mathbf{Pr}[B]$. Note that if an event A implies an event B, then this means $A \subseteq B$. Similarly, if you are trying to lower bound $\mathbf{Pr}[A]$, you can try to find B with $B \subseteq A$, and then bound $\mathbf{Pr}[B]$.
- If you are trying to upper bound $\mathbf{Pr}[A]$, you can try to lower bound $\mathbf{Pr}[\overline{A}]$ since $\mathbf{Pr}[A] = 1 - \mathbf{Pr}[\overline{A}]$. Similarly, if you are trying to lower bound $\mathbf{Pr}[A]$, you can try to upper bound $\mathbf{Pr}[\overline{A}]$.
- If you need to calculate $\mathbf{Pr}[A_1 \cap \cdots \cap A_n]$, try the chain rule. If the events are independent, then this probability is equal to the product $\mathbf{Pr}[A_1] \cdots \mathbf{Pr}[A_n]$. Note that the event "for all $i \in \{1, \ldots, n\}$, A_i" is the same as $A_1 \cap \cdots \cap A_n$.
- If you need to upper bound $\mathbf{Pr}[A_1 \cup \cdots \cup A_n]$, you can try to use the union bound. Note that the event "there exists an $i \in \{1, \ldots, n\}$ such that A_i" is the same as $A_1 \cup \cdots \cup A_n$.
- When trying to calculate $\mathbf{E}[\mathbf{X}]$, try:
  (i) directly using the definition of expectation;
  (ii) writing $\mathbf{X}$ as a sum of indicator random variables, and then using linearity of expectation.

4 Check Your Understanding

Problem. 1. Describe what a probability tree is.

2. True or false: If two events A and B are independent, then their complements $\overline{A}$ and $\overline{B}$ are also independent. (The complement of an event A is $\overline{A} = \Omega \backslash A$.)
3. True or false: If events A and B are disjoint, then they are necessarily independent.
4. True or false: For all events A, B, $\mathbf{Pr}[A \mid B] \leq \mathbf{Pr}[A]$.
5. True or false: For all events A, B, $\mathbf{Pr}[\overline{A} \mid B] = 1 - \mathbf{Pr}[A \mid B]$.
6. True or false: For all events A, B, $\mathbf{Pr}[A \mid \overline{B}] = 1 - \mathbf{Pr}[A \mid B]$.

7. True or false: Assume that every time a baby is born, there is $1/2$ chance that the baby is a boy. A couple has two children. At least one of the children is a boy. The probability that both children are boys is $1/2$.
8. What is the union bound?
9. What is the chain rule?
10. What is a random variable?
11. What is an indicator random variable?
12. What is the expectation of a random variable?
13. What is linearity of expectation?
14. When calculating the expectation of a random variable $\mathbf{X}$, the strategy of writing $\mathbf{X}$ as a sum of indicator random variables and then using linearity of expectation is quite powerful. Explain how this strategy is carried out.
15. True or false: Let $\mathbf{X}$ be a random variable. If $\mathbf{E}[\mathbf{X}] = \mu$, then $\mathbf{Pr}[\mathbf{X} = \mu] > 0$.
16. True or false: For any random variable $\mathbf{X}$, $\mathbf{E}[1/\mathbf{X}] = 1/\mathbf{E}[\mathbf{X}]$.
17. True or false: For any random variable $\mathbf{X}$, $\mathbf{Pr}[\mathbf{X} \geq \mathbf{E}[\mathbf{X}]] > 0$.
18. True or false: For any non-negative random variable $\mathbf{X}$, $\mathbf{E}[\mathbf{X}^2] \leq \mathbf{E}[\mathbf{X}]^2$.
19. True or false: For any random variable $\mathbf{X}$, $\mathbf{E}[-\mathbf{X}^3] = -\mathbf{E}[\mathbf{X}^3]$.
20. What is Markov's inequality?
21. What is a Bernoulli random variable? Give one example.
22. What is the expectation of a Bernoulli random variable, and how do you derive it?
23. What is a Binomial random variable? Give one example.
24. What is the expectation of a Binomial random variable, and how do you derive it?
25. What is a Geometric random variable? Give one example.
26. What is the expectation of a Geometric random variable (justification not required)?
27. Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables. Does the expression $\mathbf{E}[\mathbf{X} \mid \mathbf{Y}] = 0$ type-check?
28. Let A be an event. Does the expression $\mathbf{E}[A] = 0$ type-check?