Probability

Matthew Gherman and Adam Lott
Inspired by the textbook *Introduction to Probability* by Dimitri Bertsekas and John Tsitsiklis

23 February 2020

1 Optional – review of set notation and operations

Probability theory (and also pretty much all of math) is formulated in the language of **sets**. We won't worry about a formal definition, but basically a set is some collection of objects for which

- order doesn't matter, so for example the set $\{A, B, C\}$ is the same as the set $\{B, A, C\}$, and
- multiplicity doesn't matter, so for example the set $\{1, 2, 2, 3\}$ is the same as the set $\{1, 2, 3\}$.

In some cases, sets can be specified by listing all of their elements. For example, $\{1, 2, 3, 4, 5\}$ and $\{\ldots, -6, -3, 0, 3, 6, \ldots\}$ are sets. But more generally, sets are specified by a notation like

$$\begin{aligned} \{x \in \mathbb{Z} : -5 \leq x \leq 5\} &= \{-5, -4, \ldots, 4, 5\} \\ \{1/n : n \in \mathbb{N}\} &= \left\{1, \frac{1}{2}, \frac{1}{3}, \ldots\right\} \end{aligned}$$

The notation on the left side of the first equation above is read "the set of integers x such that $-5 \leq x \leq 5$".

Exercise 1.1. Write each of the following sets using the set notation described above. Note there can be many correct answers.

(a) $\{1, 3, 5, 7, \ldots\}$

(b) $\{2, 3, 5, 7, 11, \ldots\}$

(c) $[0, 4]$

We now introduce some central ideas and operations.

- If A and B are two sets such that every element of A is also an element of B, then we say A is a **subset** of B and write $A \subseteq B$. For example, $\{1, 2\} \subseteq \{1, 2, 3\}$, but $\{1, 2, 3\} \not\subseteq \{1, 2\}$.

- For a set A, the set of all elements which are *not* elements of A is called the **complement** of A and is denoted A^c. The complement of a set depends on the "universe" in which the set lives. If $A = \{2, 4, 6, \ldots\}$ is viewed as a set in $\mathbb{N}$, then $A^c = \{1, 3, 5, \ldots\}$. However if A is viewed as a set in $\mathbb{Z}$, then $A^c = \{\ldots, -2, -1, 0, 1, 3, 5, 7, \ldots\}$. The universe in question will always be clear from context.
- If A and B are two sets, then $A \setminus B$ is the set of all elements which are in A but not in B. For example, $\{1, 2, 3, 4, 5\} \setminus \{4, 5, 6, 7\} = \{1, 2, 3\}$. Note that this notation makes sense even if B is not a subset of A.
- The **union** of two sets A and B is the set of elements which are in *at least one* of A or B, and is denoted $A \cup B$. For example, $\{1, 2, 3\} \cup \{3, 4, 5\} = \{1, 2, 3, 4, 5\}$.
- The **intersection** of A and B is the set of elements which are in *both* A and B, and is denoted $A \cap B$. For example $\{1, 2, 3, 4\} \cap \{3, 4, 5, 6\} = \{3, 4\}$.

Exercise 1.2 (DeMorgan's Laws)**.** Prove that $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$. More generally, if $A_1, A_2, A_3, \ldots$ are any sets, prove that

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c \qquad \text{and}$$
$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c$$

HINT: To prove that two sets are equal, prove that each set is a subset of the other. To prove that one set is a subset of another set, start with an arbitrary element of the first set and deduce that it is a member of the second set.

Exercise 1.3. If A and B are two sets, prove that $(A \cap B)^c = (A^c \cap B) \cup (A^c \cap B^c) \cup (A \cap B^c)$.

Exercise 1.4. If $A, B_1, B_2, B_3, \ldots$ are any sets, prove that

$$A \cap \left(\bigcup_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty} (A \cap B_i)$$

and

$$A \cup \left(\bigcap_{i=1}^{\infty} B_i\right) = \bigcap_{i=1}^{\infty} (A \cup B_i)$$

2 Foundations of probability

A **probability model** (or sometimes called a **probability space**) consists of the following components:

- A set Ω of all possible **outcomes** of some experiment, called the **sample space**. A subset $E \subseteq \Omega$ (any subcollection of possible outcomes) is called an **event**. For now we will assume that Ω is a countable (discrete) set.
- A **probability measure** $\mathbb{P}$, which is a function that assigns to every event $E \subseteq \Omega$ a real number between 0 and 1 (inclusive). The number $\mathbb{P}(E)$ is called the **probability** of the event E.

In addition, to make our mathematical formalism compatible with our physical intuition of what probability means, we require the probability measure $\mathbb{P}$ to satisfy the following axioms:

(P1) **Non-negativity**: $\mathbb{P}(E) \geq 0$ for any event $E \subseteq \Omega$.

(P2) **Additivity**: If E and F are disjoint events (i.e. $E \cap F = \emptyset$), then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$. You can think of this as saying "if it's impossible for E and F to both happen, then the probability of E or F happening is the sum of the probabilities of E and F". More generally, if $E_1, E_2, \ldots$ is an infinite list of pairwise disjoint events (meaning $E_i \cap E_j = \emptyset$ for all $i \neq j$), then we assume $\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$.

(P3) **Normalization**: $\mathbb{P}(\Omega) = 1$. You can think of this as saying "It is guaranteed (probability 1) that *some* outcome will occur".

Exercise 2.1. There are many other intuitive properties that we would expect the probability measure $\mathbb{P}$ to have. One might expect that more axioms would be needed to capture these properties, but in fact all of the other properties can be derived from these three axioms. Using only the axioms above, prove the following properties of $\mathbb{P}$ and interpret them in terms of physical intuition.

(a) **Monotonicity:** If $E \subseteq F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$.

(b) **Empty set:** $\mathbb{P}(\emptyset) = 0$.

(c) **Finite additivity:** If $E_1, \ldots, E_n$ are pairwise disjoint events, then $\mathbb{P}(E_1 \cup \ldots \cup E_n) = \mathbb{P}(E_1) + \ldots + \mathbb{P}(E_n)$.

(d) **Complement rule:** For any event E, $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$. More generally, if $E \subseteq F$ then $\mathbb{P}(F \backslash E) = \mathbb{P}(F) - \mathbb{P}(E)$.

(e) **Union rule:** For any events E and F, $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$.

(f) Think of some other physically intuitive properties that probability should have, formulate them in the language of the measure $\mathbb{P}$ and events E, and prove them.

Example 1. To get a better idea of what this all means, let's look at a specific example of a probability model for a specific experiment. Consider the experiment of flipping a standard coin three times. The set of all possible outcomes is the set of all possible sequences of H and T that can appear – we write

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}.$$

Now we have to say what the probability measure $\mathbb{P}$ is. Formally, this means that we have to assign a number $\mathbb{P}(E)$ to each of the $2^8 = 256$ possible subsets $E \subseteq \Omega$, and what's more, we have to do it in such a way so that it satisfies the axioms (P1)-(P3)! This seems like an impossible task, but fortunately there is a shortcut. Because of the finite additivity property (Exercise 1c), we only have to specify the probability of each *individual* outcome, and then $\mathbb{P}(E)$ can be defined for any E by just summing up the probabilities of the individual outcomes that make up E. Then the additivity axiom will be automatically satisfied, so we just need to make sure that our definition satisfies normalization and non-negativity.

Let us define $\mathbb{P}$ by declaring that each of the eight possible outcomes is equally likely, i.e. $\mathbb{P}(HHH) = \mathbb{P}(HHT) = \cdots = \mathbb{P}(TTT) = 1/8$. (Sanity check: why did we pick the number 1/8?) This definition of $\mathbb{P}$ is the most compatible with our physical intuition of how coin flips should behave (it makes sense that any possible sequence is just as likely as any other), but there are plenty of other ways to define it in a way that still satisfies the axioms. For example, let's define a different probability measure $\mathbb{P}'$ on the same sample space by $\mathbb{P}'(HHH) = \mathbb{P}'(HHT) = \mathbb{P}'(HTH) = \mathbb{P}'(HTT) = 1/5$, $\mathbb{P}'(THH) = \mathbb{P}'(THT) = \mathbb{P}'(TTH) = \mathbb{P}'(TTT) = 1/20$.

Exercise 2.2. Verify that $\mathbb{P}'$ is a probability measure on Ω. Can you think of any situation that would make $\mathbb{P}'$ a physically intuitive assignment of probabilities? (The answer might be no)

Exercise 2.3. Often, it will be much easier to refer to events with words that describe the physical situation the event indicates, rather than formally as a subset of Ω. For example, we might say "E is the event that the third coin lands tails" to formally mean $E = \{\text{HHT, HTT, THT, TTT}\} \subseteq \Omega$.

(a) Calculate $\mathbb{P}$(the first two coins land on H).

(b) Calculate $\mathbb{P}'$(the first and third coin flips are not the same).

(c) Calculate $\mathbb{P}$(an even number of Ts appear in the sequence).

(d) Calculate $\mathbb{P}'$(the coin never lands on the same side twice in a row).

Exercise 2.4. Define a probability model (sample space Ω and probability measure $\mathbb{P}$) for each of the following experiments. You can make your probability measures physically sensible if you want, but you don't have to.

(a) Roll a standard six-sided die, and then flip a coin however many times the die says.

(b) Flip a coin over and over until you get tails for the first time.

(c) Imagine you have a stick of length 5. Break the stick at a random integer length (so it can be broken into pieces of lengths 1 and 4, 2 and 3, 3 and 2, or 4 and 1), and then take the longer piece and break it again at a random integer length.

Exercise 2.5 (CHALLENGE – The Monty Hall problem)**.** Suppose you are on a game show in which the host offers you three doors to choose from. You are told that behind two of the doors are goats and behind the third door is a new car. You pick a door at random (with each door being selected with probability 1/3), but instead of opening the door you picked, the host opens one of the *other* doors to reveal a goat. The host then asks you if you would like to switch from your original choice to the other unopened door. Should you switch? (Assume that your goal is to maximize your chances of winning the new car).

2.1 Optional – continuous probability models

In many physically relevant situations, it makes more sense to model the space of all possible outcomes as an **uncountable** (continuous) set of outcomes rather than countable (discrete). For example, if you throw a dart at a circular dartboard, the possible outcomes would be all of the points in the circle. In this section we will see how our theory can accommodate this.

The definition of a probability model $(\Omega, \mathbb{P})$ is exactly the same as it is for discrete spaces. [1] The main difference is that it is no longer possible to define $\mathbb{P}$ by just defining $\mathbb{P}(\omega)$ for each individual outcome $\omega \in \Omega$, because the axiom of additivity only applies to *countable* unions, but Ω is uncountable. The following exercises will illustrate how one can define $\mathbb{P}$ in the uncountable setting.

Exercise 2.6. Let $(\Omega, \mathbb{P})$ be any probability space (discrete or continuous). The following two results are known as the **continuity of measure**.

(a) Suppose that $E_1, E_2, \ldots$ are **increasing** events, that is $E_1 \subseteq E_2 \subseteq \ldots$, and define $E_\infty = \bigcup_{n=1}^{\infty} E_n$. Prove that $\mathbb{P}(E_\infty) = \lim_{n\to\infty} \mathbb{P}(E_n)$.

(b) Suppose that $F_1, F_2, \ldots$ are **decreasing** events, that is $F_1 \supseteq F_2 \supseteq \ldots$, and define $F_\infty = \bigcap_{n=1}^{\infty} F_n$. Prove that $\mathbb{P}(F_\infty) = \lim_{n\to\infty} \mathbb{P}(F_n)$.

Remark 1. In case you forgot the definition of a limit, here it is. If $(a_n)_{n=1}^{\infty}$ is a sequence of real numbers, we say $\lim_{n\to\infty} a_n = a$ if for each $\epsilon > 0$ there is an N such that $|a_n - a| < \epsilon$ for all $n > N$. But for the problem above, you don't need to write such a careful proof of the limit.

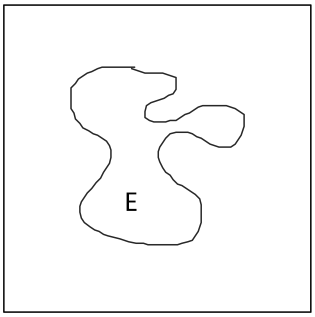
Exercise 2.7. Let Ω be the unit square $[0,1] \times [0,1]$. For all "nice" sets $E \subseteq \Omega$, let us define $\mathbb{P}(E)$ to be the area of E (for now, we ignore any questions about what "nice" or "area" mean).

[1] Part of the definition was that $\mathbb{P}(E)$ is defined for *every* event $E \subseteq \Omega$. In week 3 we will see that this is not 100% true, but for now we don't have to worry about it.

(a) Verify that $\mathbb{P}$ is a valid probability measure.

(b) Notice that if $p \in \Omega$ is any single point, then necessarily we have $\mathbb{P}(\{p\}) = 0$. Does this give you an idea of why we only allow *countable* additivity in the axioms?

(c) It's possible to come up with events $E \subseteq \Omega$ for which it's not so clear what the "area" is (see the picture below). Can you think of a way that the area of this event could be defined? (NOTE: this is just a conceptual question, you don't have to actually calculate the area)



(d) If that was too easy, try this one. Let E be the set of all points $(x, y) \in \Omega$ such that if you write out x and y in their standard decimal representation, x and y have the same digit in at least one decimal place. First, think about what the "area" of E even means. Then, calculate $\mathbb{P}(E)$. (HINT: Exercise 2.6)

Exercise 2.8. In the previous exercise we defined $\mathbb{P}$ in a geometric way without worrying too much about formal definitions or for which sets the definition made sense. Now we will be a bit more careful. Let Ω be the interval $[0, 1]$.

(a) First, for all events of the form $[a, b]$ where $0 \leq a \leq b \leq 1$, define $\mathbb{P}([a, b]) := \frac{\pi}{2} \int_a^b \sin(\pi x)\, dx$. Verify that this satisfies axioms (P1) and (P3) and show that again, $\mathbb{P}(\{p\}) = 0$ for any single point p.

(b) Define and then calculate $\mathbb{P}([0,1] \cup [3/4,1])$. More generally, if $[a_1,b_1],[a_2,b_2],\ldots$ are any pairwise disjoint intervals, then give a definition for $\mathbb{P}\left(\bigcup_{n=1}^{\infty}[a_n,b_n]\right)$. Remember that $\mathbb{P}$ must satisfy axiom (P2) to be a valid measure.

(c) Define $\mathbb{P}((a,b))$ for any open interval.

(d) How would you define $\mathbb{P}(E)$ where $E := [0,1] \setminus \mathbb{Q}$?

(e) What about really weird events? How would you try to define $\mathbb{P}(F)$ where $F = \{x \in [0,1] : \text{only prime digits appear in the decimal expansion of } x\}$?

(f) Can you think of an experiment where this might be a physically reasonable probability model to use?

These past two examples illustrate some important differences between the discrete in continuous cases. First, in continuous models, it's very common for each single point to have probability zero. Second, because of this, it's much harder to give a definition of $\mathbb{P}(E)$ for all possible events E. For now, the best we can do is define $\mathbb{P}(E)$ for a collection of "nice" events E, and then extend our definition to slightly weirder events by using additivity and continuity properties.

As a final remark, notice that it's also possible for the sample space Ω to technically be continuous, but for the measure $\mathbb{P}$ to not "see" that.

Exercise 2.9. Let $\Omega = [0,1] \times [0,1]$ and define $\mathbb{P}(E)$ for any event E by

$$\mathbb{P}(E) \;=\; \begin{cases} 1 & \text{if } (1/4,1/4) \in E \text{ and } (1/2,1/2) \in E \\ \frac{1}{2} & \text{if } (1/4,1/4) \in E \text{ or } (1/2,1/2) \in E \text{ but not both} \\ 0 & \text{if } (1/4,1/4) \notin E \text{ and } (1/2,1/2) \notin E \end{cases}$$

Show that $\mathbb{P}$ is a valid probability measure. Also convince yourself that this is the same as just assigning $\mathbb{P}(1/2,1/2) = \mathbb{P}(1/4,1/4) = 1/2$ and defining everything else according to additivity. So Ω is formally a continuous space, but according to $\mathbb{P}$ it might as well just be two points.

Exercise 2.10. There can also be a mixture of the two: let $\Omega = [0,1] \times [0,1]$ and define $\mathbb{P}$ by

$$\mathbb{P}(E) \;=\; \frac{1}{2} \cdot \text{area}(E) + \frac{1}{2} \cdot \begin{cases} 1 & \text{if } (1/2,1/2) \in E \\ 0 & \text{if } (1/2,1/2) \notin E \end{cases}$$

Show that $\mathbb{P}$ is a valid measure and explain why you can think of $(\Omega, \mathbb{P})$ as a mixture of a discrete and continuous model.

3 Conditional probability and independence

3.1 Definitions and examples

Now that we have a handle on what it means to have a probability model, it is time to introduce perhaps the most important idea in all of probability theory – the concept of **conditioning**.

Imagine the following situation. You are performing an experiment, and you have a probability model that tells you the probabilities of different events occurring. But now imagine that somehow, you are able to gain partial information about the outcome of the experiment – you still don't know exactly what happened, but maybe you know some qualitative information about the outcome. We would like our probability theory to include a way to *update* the probabilities of events in light of new information.

Example 2. Suppose you are rolling a pair of fair six-sided dice, and you will win a prize if your two dice total 10 or more. Since the dice are fair, we will use a probability model that says that each of the 36 possible outcomes is equally likely, so the probability of you winning the prize is 6/36. But now imagine that when you roll the dice, one of them falls off the table and rolls out of sight, and you can only see that the first die landed on 6. Now what is the probability that you win the prize? If the hidden die shows 4,5, or 6, you win, and if it shows 1,2, or 3, you lose, so your new probability of winning is 3/6.

Example 3. Suppose you are flipping two fair coins and you win a prize if the second coin lands on tails. But imagine that you are wearing special glasses that allow you to see only whether or not the two coins landed on the same side. Before the experiment, since each possible outcome is equally likely (probability 1/4), the two winning outcomes are HT and TT, so the probability of winning is 1/2. Now suppose you flip the coins and through your special glasses you can see that they are both showing the same side. Now the only possibilities are HH and TT, of which only TT is winning, so your probability of winning is still 1/2.

These ideas are formalized into our theory with the following definitions.

Definition 1. Let E and G be two events in some probability model $(\Omega, \mathbb{P})$, and suppose that $\mathbb{P}(G) > 0$ (G has a nonzero chance of occurring). The **conditional probability of E given G** is denoted $\mathbb{P}(E|G)$ and defined by

$$\mathbb{P}(E|G) \ := \ \frac{\mathbb{P}(E \cap G)}{\mathbb{P}(G)}.$$

We say that E is **independent** of G if $\mathbb{P}(E|G) = \mathbb{P}(E)$.

Let us take a moment to interpret these definitions physically. The conditional probability $\mathbb{P}(E|G)$ is supposed to represent the new probability of E occurring, given that we already know that G has occurred. So you can interpret the conditional probability formula as saying: we already know that G occurred, so we can shrink our "universe of possible outcomes" to only those outcomes that cause G to occur. Now we want to know the probability of E occurring in this new universe, which is the same as the probability of both E and G occurring in the old universe ($\mathbb{P}(E \cap G)$). But since we shrunk our universe to a smaller size, we need to divide by the "size" of the new universe to get the correct new probability: $\frac{\mathbb{P}(E \cap G)}{\mathbb{P}(G)}$.

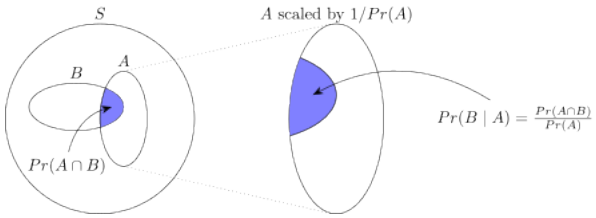
Finally, for E to be independent of G means that the probability of E occurring doesn't change once we find out that G happened. This can be interpreted as saying that whether or not G occurs has no effect on whether or not E occurs, hence the name independent.

Exercise 3.1. (a) In Example 2 above, calculate $\mathbb{P}(\text{sum is} \geq 10 \,|\, \text{first die lands on 6})$ directly from the definition.

(b) In Example 3 above, calculate $\mathbb{P}$(second coin is T | both coins are the same) directly from the definition and conclude that these two events are independent.

(c) Flip 10 fair coins and assume that all 2^{10} possible sequences are equally likely. Calculate $\mathbb{P}$(last coin is H | first nine coins are H).

Exercise 3.2. Refer to the picture below. Let $\Omega = S$ be the big circle and define $\mathbb{P}(E)$ to be the area of the event E.



Another way of thinking about conditional probability in this setting is to say that $\mathbb{P}(B|A)$ is the relative proportion of event A that is also in the event B. Convince yourself that this is saying the same thing as the formal definition from before.

Exercise 3.3 (B & T page 35)**.** Consider an experiment of rolling two four-sided dice, where each of the $4^2 = 16$ possible outcomes is equally likely. For each of the following, first use your physical intuition to guess whether or not the two events are independent. Then, use the definitions to calculate the relevant probabilities and check your guesses.

(a) $E = \{\text{first die} = 1\}$, $F = \{\text{second die} = 3\}$.

(b) $E = \{\text{maximum of the two rolls} = 2\}$, $F = \{\text{minimum of the two rolls} = 2\}$.

(c) $E = \{\text{first die} = 1\}$, $F = \{\text{sum of both dice} = 5\}$.

Remark 2. The previous two exercises teach a very important lesson: when studying probability, most of the time physical intuition can be a powerful tool, but sometimes it can lead you astray. Sometimes events that seem very dependent are actually independent and vice versa. It is usually safest to use a combination of both physical intuition and formal calculations.

Exercise 3.4. Suppose that E and G are two events with nonzero probability. Prove that if E is independent of G, then also G is independent of E. For this reason, we will usually say things like "E and G are independent events" because the order does not matter.

Exercise 3.5. Prove that if E and F are any two events with nonzero probability, then $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F|E)$. Also prove that E and F are independent if and only if $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$. Interpret these formulas physically.

Exercise 3.6. Note that our definition of independence is not defined when one of the events has probability zero (because conditional probability is not defined). But by the previous exercise, we can extend the definition to say that events E and F are independent if $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$. With this extended definition, prove that if $\mathbb{P}(E) = 0$, then E is independent of any other event. Also prove that if $\mathbb{P}(E) = 1$, then E is independent of any other event. Interpret.

Exercise 3.7. Let $(\Omega, \mathbb{P})$ be a probability space and let G be any event with $\mathbb{P}(G) > 0$. Define a new probability measure $\mathbb{P}'$ on Ω by $\mathbb{P}'(E) := \mathbb{P}(E|G)$ for any event E. Show that $\mathbb{P}'$ is a valid probability measure satisfying $\mathbb{P}'(G) = 1$. This new measure is sometimes called the **conditional measure given** G.

3.2 The tower property

One of the most useful properties of conditional probability is that it allows us to calculate probabilities of more complicated events that it would be difficult to analyze directly. Let's prove a theorem called the **tower property** (or sometimes the **law of total probability**) and then see some examples of its applications.

Theorem 1 (Tower property of conditional probability)**.** *Suppose that $E_1, E_2, \ldots, E_n$ are events that partition the sample space Ω (that is, the E_i are pairwise disjoint and $E_1 \cup \ldots \cup E_n = \Omega$). Suppose that each $\mathbb{P}(E_i) > 0$. Then for any event F,*

$$\mathbb{P}(F) \;=\; \mathbb{P}(F|E_1)\mathbb{P}(E_1) + \mathbb{P}(F|E_2)\mathbb{P}(E_2) + \ldots + \mathbb{P}(F|E_n)\mathbb{P}(E_n).$$

We can interpret this theorem in the following way. We want to know the probability of an event F, but the event F is too complicated to analyze directly. Instead, we can break up the sample space into several possible cases (the events E_i). If we assume that we are in one of our cases, then the event F becomes much simpler to analyze. Therefore we can calculate the conditional probability of F for each of our possible cases, and to get the total probability of F we just average those conditional probabilities according to the probabilities of each of the cases.

After proving the theorem, we will see some examples.

Exercise 3.8. Prove the tower property. Also state and prove a version of it in the case of an infinite partition: pairwise disjoint events $E_1, E_2, \ldots$ such that $\bigcup_{i=1}^{\infty} E_i = \Omega$.

Exercise 3.9 (B & T page 30)**.** You roll a fair four-sided die. If the result is 1 or 2, you roll again, but otherwise you stop. Calculate the probability that the sum of all your rolls is at least 4.

Exercise 3.10 (B & T page 58)**.** You have M jars, each of which contains 3 white balls and 2 black balls. A ball is (uniformly) randomly chosen from the first jar and transferred to the second jar. Then a ball is randomly chosen from the second jar and transferred to the third jar. This process continues until a ball is transferred into the last jar. Finally, a ball is chosen uniformly at random from the Mth jar. Find the probability that the last ball chosen is white.

Exercise 3.11 (B & T page 57)**.** Boris is about to play a two-game chess match with an opponent, and wants to find the strategy that maximizes his winning chances. Each game ends with either a win by one of the players, or a draw. If the score is tied at the end of the two games, the match goes into sudden-death mode, and the players continue to play until the first time one of them wins a game (and the match). Boris has two playing styles, *timid* and *bold*, and he can choose one of the two at will in each game, no matter what style he chose in previous games.

With timid play, he draws with probability $p_d > 0$, and he loses with probability $1 - p_d$. With bold play, he wins with probability $p_w > 0$, and he loses with probability $1 - p_w$. Boris will always play bold during sudden death, but may switch style between games 1 and 2.

(a) Find the probability that Boris wins the match if he plays bold in both games 1 and 2.

(b) Find the probability that Boris wins the match if he plays timid in both games 1 and 2.

(c) Find the probability that Boris wins the match if he adopts the strategy of playing timid whenever he is ahead in the score, and bold if he is tied or losing.

(d) Assume that $p_w < 1/2$, so Boris is the worse player, regardless of the playing style he adopts. Show that with the strategy in part (c) above, and depending on the values of p_w and p_d, the probability that Boris wins the match can still be greater than $1/2$. How do you explain this advantage?

3.3 Bayes' theorem

The law of total probability is often used in conjunction with Bayes' Rule, a way of relating conditional probabilities of the form $\mathbb{P}(A|B)$ with conditional probabilities of the form $\mathbb{P}(B|A)$.

Theorem 2 (Bayes' Rule)**.** *Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbb{P}(A_i) > 0$ for all i. Then, for any event B such that $\mathbb{P}(B) > 0$, we have*

$$\begin{aligned}\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(A_1)\mathbb{P}(B|A_1) + \cdots + \mathbb{P}(A_n)\mathbb{P}(B|A_n)}.\end{aligned}$$

Bayes' Rule is most often used for inference. We may have identify many causes, $A_1, \ldots, A_n$, of some effect B, and we would like to figure out which cause is most likely. The probability $\mathbb{P}(B|A_i)$ is the chance the effect will occur given a certain cause. Bayes' Rule allows us to evaluate $\mathbb{P}(A_i|B)$, the probability that the cause A_i is present given the effect.

Exercise 3.12. (a) Prove the first equality of Bayes' Rule using the definition of conditional probability.

(b) Verify the second equality of Bayes' Rule using the tower property.

Exercise 3.13. Let $A = \{\text{an aircraft is present}\}$ and $B = \{\text{the radar generates an alarm}\}$. We are given $\mathbb{P}(A) = 0.05$, $\mathbb{P}(B|A) = 0.99$, and $\mathbb{P}(B|A^c) = 0.1$. Apply Bayes' Rule to find $\mathbb{P}(A|B)$ for $A_1 = A$ and $A_2 = A^c$. Interpret the results.

Exercise 3.14 (The False-Positive Puzzle)**.** A test for a certain rare disease is assumed to correct 95% of the time. In other words, if a person has the disease, the test results are positive with a probability of 0.95. If a person does not have the disease, the test results are negative with a probability of 0.95. A random person drawn from a certain population has a probability of 0.001 of having the disease. Given that the person just tested positive, what is the probability of having the disease? Interpret the result.

3.4 Optional: additional topics in independence

3.4.1 Pairwise vs total independence

We can also define what it means for a group of more than two events to be independent, but now there are two different definitions to deal with.

Definition 2. A family of events $E_1, E_2, \ldots, E_n$ are said to be **pairwiseindependent** if E_i and E_j are independent for each pair $i \neq j$. The family is said to be **totally independent** if for *any* subcollection of the events, the probability of their intersection is equal to the product of their probabilities. More formally, we can write this as

$$\mathbb{P}\left(\bigcap_{i\in S} E_i\right) = \prod_{i\in S} \mathbb{P}(E_i) \quad \text{for any } S \subseteq \{1, 2, \ldots, n\}.$$

Exercise 3.15. Prove that total independence implies pairwise independence.

Exercise 3.16. Give an example of a family of events which is pairwise independent but not totally independent.

Exercise 3.17. Give an example of events A, B, C such that $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ but the family is not even pairwise independent.

3.4.2 Conditional independence

Definition 3. Let A, B, C be events with positive probability. We say that A and B are **conditionally independent** given C if $\mathbb{P}(A|C)\mathbb{P}(B|C) = \mathbb{P}(A \cap B|C)$.

Exercise 3.18. Prove that the above definition of conditional independence is equivalent to the statement that $\mathbb{P}(A|B \cap C) = \mathbb{P}(A|C)$.

Two comments about the notion of conditional independence:
(1) You can think about the statement "A and B are conditionally independent given C" as saying: A and B might have some dependence, but the event C already includes all the information that B can tell you about A, so if you know C happened, then knowing B also happened doesn't give you any *more* information about the occurrence of A.

(2) Another way to say that A and B are conditionally independent given C is to say that A and B are independent events in the new probability space $(\Omega, \mathbb{P}')$ where $\mathbb{P}'$ is the conditional measure given C (refer back to exercise 3.7).

Exercise 3.19. Find an example of three events A, B, C such that

(a) A and B are not independent, but they are conditionally independent given C.

(b) A and B are independent, but they are *not* conditionally independent given C. Think about how to interpret this case physically.

Exercise 3.20. State and prove a "conditional tower property" (this question doesn't really have anything to do with independence).
HINT: There is a long way using direct calculation or a short way using the conditional measure (exercise 3.7).

Exercise 3.21. Suppose events A, B, C are totally independent. Prove that A and B are conditionally independent given C.

3.5 Optional: conditional probability in continuous probability spaces

If $(\Omega, \mathbb{P})$ is a continuous probability space, the definitions of conditional probability and independence are exactly the same.

Exercise 3.22. Let $\Omega = [0,1] \times [0,1]$ and $\mathbb{P}$ be area. Denote an outcome in Ω by (x, y). Calculate

(a) $\mathbb{P}(x > y \,|\, y < 1/2)$

(b) $\mathbb{P}(x + y < 1/2 \,|\, x < y^2)$

The tower property is also still true in the continuous setting. So why is this even an extra section? The main difficulty in the continuous setting is the presence of probability zero events. A big limitation in the definition of conditional probability is that you are only allowed to condition on events of positive probability. In the discrete setting this is not so bad because usually your model is set up so that each individual outcome has positive probability and therefore the only probability zero event is the empty set. But whenever we want to apply the tower property in the continuous setting, this causes a huge problem. Consider the following example.

First, pick a starting point x uniformly at random from the interval $[0, 1]$. Then, pick a jumping distance y uniformly at random from the interval $[0, x]$ and jump from x to $x + y$. What is the probability that your ending point is greater than 1?

This probability is hard to figure out directly because the size of your jump depends on your random starting point, but it's a perfect situation to apply the tower property. We would want to write something like

$$\mathbb{P}(\text{ending point } > 1) \;=\; \sum_{x \in [0,1]} \mathbb{P}(\text{starting point } = x)\mathbb{P}(\text{jump length } > 1 - x \,|\, \text{starting point} = x).$$

But there are lots of things wrong with this. First, the sum is over an uncountable set so it's not even defined. Second, the event $\{\text{starting point} = x\}$ is probability zero, so the conditional probability given it is not defined either. Third, even if it were defined, the presence of the $\mathbb{P}\{\text{starting point} = x\}$ term in the sum would make every term equal to zero anyway!

The first problem is not so bad – as you might guess, we can just use an integral instead of a sum. But the second and third problems are very serious and to solve them we need to develop much more machinery. We might be able to touch on this more in the second and third week, so stay tuned.

Probability II

Matthew Gherman and Adam Lott
Inspired by the textbook *Introduction to Probability* by Dimitri Bertsekas and John Tsitsiklis

1 March 2020

4 Random variables

4.1 Basics

In many random experiments, the outcomes themselves are not literally numbers, but we want to associate some numerical information to the outcome because numbers are things that we know how to analyze. Also, depending on what particular aspects of the outcomes we care about, we might want to associate many different pieces of numerical information to them. For example, suppose we flip a coin 100 times. Formally, the outcomes are strings of Hs and Ts of length 100, but there are plenty of numerical data we might care about. For example, maybe we care about the total number of Hs that were flipped, or the length of the longest string of consecutive Ts, or the number of times the particular sequence HTTHT occurred. A **random variable** is a way of assigning numerical information to the outcome of a random experiment. We describe this formally by saying that a random variable X is any function $X : \Omega \to \mathbb{R}$.

Exercise 4.1. Think of a few different random experiments and come up with some examples of random variables for those experiments that you might care about.

Definition 4. If X is a random variable on the probability space $(\Omega, \mathbb{P})$, then we can define what is called the **distribution** of X (also sometimes called the **probability mass function** of X or the **law** of X). The distribution of X, denoted dist_X, can be thought of as a "function" [1] that assigns to each $a \in \mathbb{R}$ the number $\text{dist}_X(a) := \mathbb{P}\{\omega \in \Omega : X(\omega) = a\}$ (this second expression will also be written as $\mathbb{P}(X = a)$ for short). We can also extend the definition to apply to any subset $B \subseteq \mathbb{R}$, meaning that we will define $\text{dist}_X(B) := \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\} = \mathbb{P}(X \in B)$.

Exercise 4.2. (a) Flip three fair coins and let X be the total number of tails. Find $\text{dist}_X(2)$.

(b) Roll two fair six-sided dice and let X be the product of the two rolls. Find $\text{dist}_X(4)$.

[1] You might be wondering why the word "function" appears in quotes, since what we have defined as the distribution is literally a function $\mathbb{R} \to \mathbb{R}$. The answer is that (a) this definition is no longer appropriate when Ω is a continuous probability space, and (b) we want to think of the distribution as something that assigns a numerical value to every *set* in $\mathbb{R}$, not just every point in $\mathbb{R}$. More on this topic coming in week 3.

(c) Roll two fair six-sided dice and let Y be the sum of the two rolls. Find dist_Y. (This means that you should find all possible values for which dist_Y is nonzero, and calculate the dist_Y at those places.)

(d) (B&T page 119) The UCLA soccer team has two games scheduled for one weekend. It has a 0.4 probability of not losing the first game, and a 0.7 probability of not losing the second game, independent of the first. If it does not lose a particular game, the team is equally likely to win or tie, independent of what happens in the other game. The UCLA team will receive two points for a win, one for a tie, and zero for a loss. Let Z be the total number of points the team earns in the weekend. Find dist_Z.

Exercise 4.3 (B & T page 120)**.** You just rented a large house and the realtor gave you five keys, one for each of the five doors of the house. Unfortunately, all keys look identical, so to open the front door, you try them at random.

(a) Find the distribution of the number of trials you need to open the front door if after each unsuccessful trial, you mark the corresponding key so that you never try it again.

(b) Same question, but on each trial you are equally likely to choose any key.

Exercise 4.4. Flip a fair coin 10 times and let X be the total number of heads and Y be the total number of tails. Show that X and Y are not the same random variable but they do have the same distribution.

Exercise 4.5. For each of the following, find an example of a probability model $(\Omega, \mathbb{P})$ and a random variable X which has the given distribution.

(a) $\text{dist}_X(0) = 1/9$, $\text{dist}_X(1) = 4/9$, $\text{dist}_X(2) = 4/9$ (NOTE: there is always an obvious answer – let $\Omega = \{0, 1, 2\}, \mathbb{P}(0) = 1/9, \mathbb{P}(1) = \mathbb{P}(2) = 4/9$, and $X(\omega) = \omega$. Try to find an example that corresponds to a physical situation.)

(b) $\text{dist}_X(k) = (1/2)^k$ for all $k = 1, 2, 3, \ldots$

(c) $\text{dist}_X(k) = \binom{10}{k}(1/4)^k(3/4)^{10-k}$ for all $k = 0, 1, 2, \ldots, 10$.

Remark 3. The previous two exercises point out an important concept. Sometimes random variables which correspond to very different sources of randomness (the underlying experiment) can produce the same distribution. In many cases, the *distribution itself is much more important than the actual source of randomness and choice of variable that produced it.* If this confuses you, don't worry, we'll come back to it later.

Exercise 4.6. Let X be any random variable. Prove that

$$\sum_{a \in \mathbb{R}} \text{dist}_X(a) \;=\; 1.$$

(Recall that since Ω is a countable set, X can take only countably many values. So you should interpret the notation $\sum_{a \in \mathbb{R}} f(a)$ as "sum over those values a for which $f(a)$ is not zero".)

4.2 Expected value

Given a random variable X, its distribution dist_X gives us complete information about what values X can take and how likely each value is. But sometimes it is easier to deal with "coarser" information about X that can be described with a single number. The notion of the **expected value** of a random variable is a way of describing the "average value" that you expect a random variable to take.

As an example, consider spinning a spinner that lands on 1 with probability 1/2, 2 with probability 1/4, and 3 with probability 1/4, and let X be the number the spinner lands on. What does the "average value" of X mean? One way to interpret this is to repeat the same experiment lots and lots of times, and take the average of all the results you get. Suppose we spin the same spinner N times, where N is very large. Then we would expect the spinner to land on 1 about $N/2$ of those times and on 2 and 3 each about $N/4$ times. So the average value of all the results would be about

$$\frac{1 \cdot (N/2) + 2 \cdot (N/4) + 3 \cdot (N/4)}{N} \;=\; 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4}.$$

This says that the average value we expect to get from X over many repeated trials can just be written as a weighted average of the possible values of X, where the weights are given by the probabilities of each value appearing. Let's now turn this idea into a definition.

Definition 5. Let X be a random variable. The **expected value** (or **expectation**) of X is defined to be the number

$$\mathbb{E}[X] \;:=\; \sum_{a \in \mathbb{R}} a \cdot \mathrm{dist}_X(a).$$

Exercise 4.7. (a) Calculate the expectation of a single roll of a fair six-sided die.

(b) Flip three fair coins independently and let Y be the total numer of heads. Calculate $\mathbb{E}[Y]$.

(c) (B & T page 122) Fix positive integers $a \leq b$. Let Z be a random variable that takes as values, with equal probability, the powers of 2 in the interval $[2^a, 2^b]$. Calculate $\mathbb{E}[Z]$.

(d) (CHALLENGE) Suppose you have a biased coin that lands on heads with probability p and tails with probability $1 - p$. Flip the coin repeatedly (each flip is independent) until getting heads for the first time, and let X be the total number of flips required. Calculate $\mathbb{E}[X]$ and interpret the result.

Exercise 4.8 (B & T page 123)**.** A prize is placed uniformly at random in one of N boxes, numbered 1 through N. You search for the prize by asking yes-no questions. Find the expected number of questions until you are sure of the location of the prize under each of the following strategies.

(a) Enumeration – you ask questions of the form "is it in box k?"

(b) Bisection – you ask questions of the form "is it in a box numbered less than or equal to k?"

Exercise 4.9 (B & T page 91)**.** You are on a quiz show and you are given two questions to answer, and you must decide which one to answer first. Question 1 will be answered correctly with probability 0.8 and a correct answer yields a prize of \$100. Question 2 will be answered correctly with probability 0.5 and a correct answer yields a prize of \$200. If you answer the first question wrong, you lose and don't get to attempt the second question. Assume your answers to the questions are independent of each other. Determine the best strategy if:

(a) Your goal is to maximize your expected earnings.

(b) Your goal is to maximize your chances of winning something.

Exercise 4.10 (The St. Petersburg paradox)**.** Flip a fair coin repeatedly until getting heads for the first time. If the first heads appears on the nth flip, you win 2^n dollars.

(a) Calculate your expected winnings.

(b) Suppose there is an entry fee for this game and you are allowed to play as many times as you want. Based on your answer to part (a), would you be willing to pay \$10/game to play? What about \$10,000,000,000/game? Explain why the answer dictated by part (a) is not a very good practical strategy.

Exercise 4.11. Let X be a random variable that only takes positive integer values. Prove that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

4.3 Functions of random variables

Given a random variable X, we can apply some transformation $f : \mathbb{R} \to \mathbb{R}$ to it to create a new random variable $f(X)$. For example, if X is the number of heads seen in a sequence of 10 coin flips, then X^2, $2X + 5$, and $\log(|\sin X|)$ are all just other random variables on the same probability space. In this section we will see that if you know information about X, then it's not too hard to determine information about any $f(X)$ also.

Exercise 4.12 (B & T page 122)**.** Let X be a random variable which takes the values 0 through 9 each with probability 1/10. Find the distributions of the random variables

(a) $X \bmod 3$

(b) $5 \bmod (X + 1)$

Exercise 4.13. Let X be any random variable and $f : \mathbb{R} \to \mathbb{R}$ be any transformation. Prove that

$$\mathbb{E}[f(X)] = \sum_{a \in \mathbb{R}} f(a) \operatorname{dist}_X(a)$$

and interpret this formula.
NOTE: this formula is *extremely* useful because it says that as long as you know the distribution of X, you can calculate the expectation of any function of X without having to recalculate the new distribution.

We will now introduce another very important statistic associated to a random variable, called the **variance**. While the purpose of the expected value is to determine the average "size" of X, the purpose of the variance is to quantify how "spread out" the distribution of X is by looking at how far away X is from its expected value.

Definition 6. The **variance** of a random variable X is defined as

$$\operatorname{var}(X) \;:=\; \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Exercise 4.14. Let X take the values ± 1 each with probability $1/2$, and let Y take the values ± 100 each with probability $1/2$. Verify that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ but that $\operatorname{var}(Y)$ is much larger than $\operatorname{var}(X)$. This example shows how variance captures the notion of how spread out the distribution of a random variable is.

Exercise 4.15. Prove that $\operatorname{var}(X) \geq 0$ and $\operatorname{var}(X) = 0$ if and only if X is deterministic, that is there is some $a \in \mathbb{R}$ so that $\mathbb{P}(X = a) = 1$.

Exercise 4.16. Prove the alternate formula

$$\operatorname{var}(X) \;=\; \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Use this to deduce the interesting inequality $\mathbb{E}\left[X^2\right] \geq \mathbb{E}[X]^2$. NOTE: in practice, this formula for the variance is usually much easier to calculate than the original defintion.

Exercise 4.17. Let X be any random variable and let $c > 0$ be a number. Calculate the expectation and variance of the random variables cX and $X + c$ in terms of $\mathbb{E}[X]$ and $\operatorname{var}(X)$.

Exercise 4.18 (OPTIONAL)**.** Let Y take the values $k = 1, 2, 3, \ldots$ with probability Z/k^3 (where Z is just the normalization constant). Show that $\mathbb{E}[Y]$ is finite but $\operatorname{var}(Y)$ is infinite.

5 Collections of many random variables

5.1 Joint distributions

Often in an experiment, there are many different random variables that we care about, and we also want to keep track of how they relate to each other. In this section we will learn about how to extend the theory from above to study many random variables simultaneously.

Definition 7. Let X and Y be two random variables on the same probability space. We can consider the pair (X, Y) to be a kind of "random variable" that takes values in $\mathbb{R}^2$ instead of in $\mathbb{R}$. We then can define the **joint distribution** of X and Y to be the function $\text{dist}_{X,Y}$ which assigns to each pair $(a, b) \in \mathbb{R}^2$ the probability $\text{dist}_{X,Y}(a, b) = \mathbb{P}(X = a \text{ and } Y = b)$.

Exercise 5.1. Flip three fair coins. Let X be the total number of heads and Y be the length of the longest string of consecutive tails. Calculate $\text{dist}_{X,Y}$. It may be easiest to organize this information in a table.

Exercise 5.2. Given the joint distribution of two random variables $\text{dist}_{X,Y}$, show that we can recover the individual distributions of X and Y by the formulas

$$\text{dist}_X(a) \;=\; \sum_{b\in\mathbb{R}} \text{dist}_{X,Y}(a, b), \qquad \text{dist}_Y(b) \;=\; \sum_{a\in\mathbb{R}} \text{dist}_{X,Y}(a, b).$$

The distributions of the individual random variables are sometimes called the **marginal** distributions of the joint distribution.

We can also perform transformations on pairs of random variables in the same way that we did on single random variables. If X, Y are two random variables and $g : \mathbb{R}^2 \to \mathbb{R}$ is any function, then $g(X, Y)$ is a new random variable. For example, $X^2 + Y^3$ or $\frac{X-Y}{2}$.

Exercise 5.3. As you might expect (refer back to exercise 4.13), we can calculate the expected value of any function of X and Y by just knowing the joint distribution. Prove the formula

$$\mathbb{E}\left[g(X, Y)\right] \;=\; \sum_{a,b\in\mathbb{R}} g(a, b)\,\text{dist}_{X,Y}(a, b).$$

Exercise 5.4. Let X and Y be any two random variables on the same probability space. Prove that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ and give an example to show that $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ in general. The first property is known as the **linearity of expectation**.

5.2 Independence

Just as we had notions of conditioning and independence for pairs of events, we also have similar ideas for pairs of random variables.

Definition 8. Two random variables X and Y are **independent** if for every $a \in \mathbb{R}$ and $b \in \mathbb{R}$, the events $\{X = a\}$ and $\{Y = b\}$ are independent events.

Exercise 5.5. Prove that X and Y are independent if and only if $\text{dist}_{X,Y}(a,b) = \text{dist}_X(a) \cdot \text{dist}_Y(b)$ for all a, b.

Exercise 5.6. Prove that if X and Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Give an example to show that even if X and Y are not independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ is still possible. Random variables X, Y for which that formula holds are called **uncorrelated**. Also prove that if X and Y are independent, then $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$, and give an example to show that this is not necessarily true if X and Y are not independent.

Exercise 5.7. So far we have defined everything only for two random variables X and Y. But everything we have done can be defined in a very similar way for any number of random variables $X_1, X_2, \ldots, X_n$.

(a) Prove that if $X_1, \ldots, X_n$ are independent random variables each with distribution $\text{Ber}(p)$, then $Y = X_1 + \ldots + X_n$ has distribution $\text{Binom}(n,p)$.

(b) Use the above fact to recalculate the expectation and variance of $\text{Binom}(n,p)$ in a much easier way.

Exercise 5.8 (OPTIONAL)**.** Prove that if $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$ and X and Y are independent, then $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$.

6 Common distributions

6.1 Bernoulli distribution

Definition 9. Let p be a fixed number between 0 and 1. We say that a random variable X has the **Bernoulli distribution** with parameter p if

$$\mathbb{P}(X = 1) \;=\; p, \qquad \mathbb{P}(X = 0) \;=\; 1 - p.$$

This is often notated $X \sim \text{Ber}(p)$. This definition and notation are reflective of the philosophy mentioned above, that the distribution is more important than the random variable itself. Formally, X is defined on some probability space Ω, but we don't care if Ω has two, two billion, or infinity elements in it, and we don't care what the elements of Ω look like. All that matters is that when we look at the values of the random variable X, it comes out to 1 with probability p and 0 with probability $1 - p$.

Exercise 6.1. What physical situation does a Bernoulli distribution correspond to?

Exercise 6.2. Calculate the expectation and variance of the distribution $\text{Ber}(p)$ and interpret the expectation.

6.2 Geometric distribution

Definition 10. Let p be a number between 0 and 1. The **geometric distribution** with parameter p, denoted $\text{Geom}(p)$, is the distribution of a random variable X such that

$$\mathbb{P}(X = k) \;=\; p(1-p)^{k-1} \qquad k = 1, 2, 3, \ldots$$

Exercise 6.3. Verify that this is a valid probability distribution.

Exercise 6.4. What physical situation does a geometric distribution correspond to?

Exercise 6.5 (CHALLENGE)**.** Calculate the expectation and variance of the distribution Geom(p) and interpret the expectation.

6.3 Binomial distribution

Definition 11. Let n be a positive integer and p be a number between 0 and 1. The **binomial distribution** with parameters n and p, denoted Binom(n, p), is the distribution of a random variable X such that

$$\mathbb{P}(X = k) \;=\; \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n.$$

Exercise 6.6. Verify that this is a valid probability distribution.

Exercise 6.7. What physical situation does a binomial distribution correspond to?

Exercise 6.8 (CHALLENGE)**.** Calculate the expectation and variance of the distribution Binom(n, p) and interpret the expectation.

6.4 Optional – Poisson distribution

Definition 12. Let λ be a positive number. The **Poisson distribution** with parameter λ, denoted Poi(λ), is the distribution of a random variable X such that

$$\mathbb{P}(X = k) \;=\; e^{-\lambda} \cdot \frac{\lambda^k}{k!} \qquad k = 0, 1, 2, \ldots.$$

Exercise 6.9. Verify that this is a valid probability distribution.

Exercise 6.10. Calculate the expectation and variance of the distribution $\mathrm{Poi}(\lambda)$.

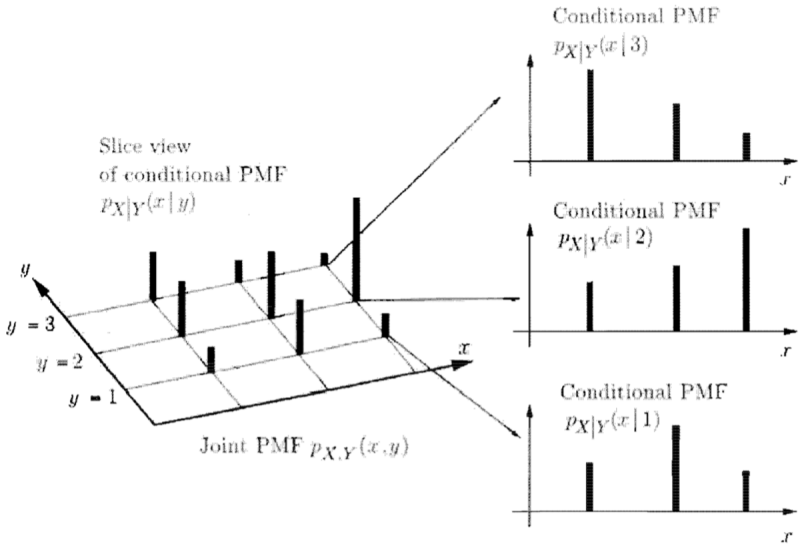
7 Conditioning

Definition 13. Let X and Y be two random variables defined on the same probability space. Let $b \in \mathbb{R}$ be a fixed value for which $\mathbb{P}(Y = b)$ is positive. We define the **conditional distribution of X given the event $\{Y = b\}$** to be the function $\text{dist}_{X|Y}(\cdot|b)$ [2] which assigns to each $a \in \mathbb{R}$ the number

$$\text{dist}_{X|Y}(a|b) \;=\; \frac{\mathbb{P}(X = a, Y = b)}{\mathbb{P}(Y = b)} \;=\; \frac{\text{dist}_{X,Y}(a, b)}{\text{dist}_Y(b)}.$$

You should think of the conditional distribution as saying: "Suppose we already know that Y took the value b. How does this change the likelihoods of X taking all of its different values?"

The picture below may help you visualize the situation.



Now that we have a notion of conditional distribution, we can define conditional expectation. The **conditional expectation of X given $\{Y = b\}$** is defined by

$$\mathbb{E}\,[X|Y = b] \;=\; \sum_a a \cdot \text{dist}_{X|Y}(a|b).$$

Exercise 7.1. Roll two fair six-sided dice. Let X be the sum of the rolls and let Y be the result of the first roll. Calculate $\mathbb{E}\,[X|Y = b]$ for each $b = 1, 2, \ldots, 6$.

Exercise 7.2. Let X and Y be two random variables. Prove the **tower property of conditional expectation**:

$$\mathbb{E}\,[X] \;=\; \sum_b \mathbb{E}\,[X|Y = b] \cdot \text{dist}_Y(b)$$

[2] The "$\cdot$" is just a symbol that tells you where the argument of the function goes.

and interpret the formula.

Exercise 7.3. Roll a die, then flip a coin as many times as the die says. What is the expected total number of heads that appear?

Exercise 7.4. A professor is holding office hours for her probability class. Students that come to office hours can have easy questions or hard questions. It takes the professor 2 minutes to answer hard questions and 1 minute to answer easy questions. Assume that all students are independent of each other and that each student shows up with H hard questions and E easy questions, where $H \sim \text{Ber}(\frac{3}{4})$ and $E \sim \text{Ber}(\frac{1}{4})$ are independent of each other. Also, the number of students that come to office hours is random with distribution Poi(10). Assume that the professor can only talk to one student at a time. Find the expected number of minutes office hours will last.

Probability supplement

Matthew Gherman and Adam Lott

8 March 2020

8 Miscellaneous puzzles

Exercise 8.1. You are given two envelopes and told that inside each envelope is a card with a positive integer written on it, but you are not given any information about what the two numbers are. Your goal is to pick the envelope with the bigger number. You are allowed to look inside one of the envelopes and then decide if you want to keep that one or switch to the other one. Is there any strategy that will allow you to win with greater than 1/2 probability? (A "strategy" is defined as any process that takes the number in the first envelope as an input, performs any number of random or deterministic experiments, and then returns a decision to stay or switch)

Exercise 8.2. There are 100 people about to board an airplane with 100 seats. Each person is assigned one unique seat. The first passenger boards the plane but then realizes that he lost his ticket, so he sits in a seat uniformly at random. The passengers on this plane are all a little bit shy, so each subsequent passenger will sit in his/her own assigned seat if it is open, but if it is taken he/she will choose a seat uniformly at random from the seats still available. What is the probability that the final passenger sits in his assigned seat?

Exercise 8.3 (B & T page 128)**.** A coin that has probability of heads equal to p is tossed successively and independently until a head comes twice in a row or a tail comes twice in a row. Find the expected value of the number of tosses.

Exercise 8.4 (B & T page 128)**.** A spider and a fly move along a straight line. At each second, the fly moves a unit step to the right or to the left with equal probability p, and stays where it is with probability $1 - 2p$. The

spider always takes a unit step in the direction of the fly. The spider and the fly start D units apart, where D is a random number with distribution Geom(1/2). If the spider lands on top of the fly, it's the end. What is the expected value of the time it takes for this to happen?

9 Non-measurable sets

9.1 Motivation

You may be familiar with the following surprising fact:

Theorem 3 (Banach-Tarski paradox)**.** *Let B be a solid sphere in 3 dimensions. Then it is possible to partition B into finitely many pieces such that if the pieces are rotated and translated, they can be reassembled to form two identical disjoint spheres both congruent to B.*

Let's think about this in terms of probability spaces. Let $\Omega = B$ be the solid sphere, and let $\mathbb{P}$ be normalized volume, so $\mathbb{P}(E) = \text{vol}(E)/\text{vol}(B)$ for any $E \subseteq B$.

Exercise 9.1. Explain why the Banach-Tarski paradox and our axiomatic definition of probability spaces contradict each other.

9.2 Construction of a non-measurable set

Giving a complete construction of the Banach-Tarski paradox is too hard, but in this section we will describe an easier construction that illustrates the same point.

Define a probability space $\Omega = [0, 2]$ and $\mathbb{P}$ = normalized length, that is for any event E, $\mathbb{P}(E) = (1/2)\cdot\text{length}(E)$.

Exercise 9.2. Define a relation $\sim$ on $[0, 1]$ by declaring $x \sim y$ if $x - y$ is a rational number. Prove that $\sim$ is an equivalence relation.

Exercise 9.3. Define our event E to consist of one representative from each equivalence class of $\sim$ (in case you are interested, this step requires the Axiom of Choice). Prove that

(a) For any two distinct rational numbers $p, q \in [0, 1]$, $E + p$ and $E + q$ are disjoint. $E + p$ notes the translate of E by p: $E + p = \{x + p : x \in E\}$.

(b) $\bigcup_{p\in\mathbb{Q}\cap[0,1]}(E + q) \subseteq [0, 2]$.

Exercise 9.4. Notice that since $\mathbb{P}$ is length, it must be the case that $\mathbb{P}(E) = \mathbb{P}(E + q)$ for any translate $E + q$. Use this to prove that $\mathbb{P}(E) = 0$.

Exercise 9.5. Prove also that $\bigcup_{p \in \mathbb{Q} \cap [0,1]} (E + q) \supseteq [0, 1]$ and deduce from this a contradiction.

As the above construction shows, there exist situations in which it is simply not possible for $\mathbb{P}(E)$ to be defined for every single $E \subseteq \Omega$ if we want $\mathbb{P}$ to satisfy the axioms of a probability measure. The question of how to fix this problem belongs to a field called *measure theory*, which we won't get into yet.