Chapter 2. Discrete Probability

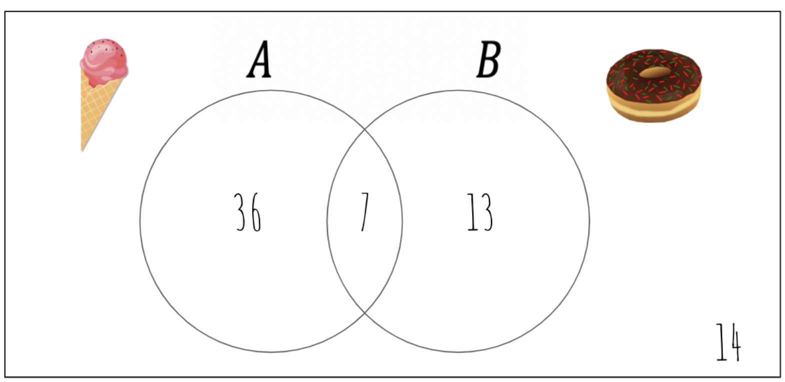
2.2: Conditional Probability

Slides (Google Drive) | Alex Tsun | Video (YouTube)

2.2.1 Conditional Probability

Sometimes we would like to incorporate new information into our probability. For example, you may be feeling symptoms of some disease, and so you take a test to see whether you have it or not. Let D be the event you have a disease, and T be the event you test positive (T^C is the event you test negative). It could be that $\mathbb{P}(D) = 0.01$ (1% chance of having the disease without knowing anything). But how can we update this probability *given* that we tested positive (or negative)? This will be written as $\mathbb{P}(D \mid T)$ or $\mathbb{P}\left(D \mid T^C\right)$ respectively. You would think $\mathbb{P}(D \mid T) > \mathbb{P}(D)$ since you're more likely to have the disease once you test positive, and $\mathbb{P}\left(D \mid T^C\right) < \mathbb{P}(D)$ since you're less likely to have the disease once you test negative. These are called conditional probabilities - they are the probability of an event, given that you know some other event occurred. Is there a formula for updating $\mathbb{P}(D)$ given new information? Yes!

Let's go back to the example of students in CSE312 liking donuts and ice cream. Recall we defined event A as liking ice cream and event B as liking donuts. Then, remember we had 36 students that only like ice cream ($A \cap B^C$), 7 students that like donuts and ice cream ($A \cap B$), and 13 students that only like donuts ($B \cap A^C$). Let's also say that we have 14 students that don't like either ($A^C \cap B^C$). That leaves us with the following picture, which makes up the whole sample space:



Now, what if we asked the question, what's the probability that someone likes ice cream, **given** that we know they like donuts? We can approach this with the knowledge that 20 of the students like donuts (13 who don't like ice cream and 7 who do). What this question is getting at, is: given the knowledge that someone likes donuts, what is the chance that they also like ice cream? Well, 7 of the 20 who like donuts like ice cream, so we are left with the probability $\frac{7}{20}$. We write this as $\mathbb{P}(A \mid B)$ (read the "probability of A given B") and in this case we have the following:

$$
\begin{aligned}
\mathbb{P}(A \mid B) &= \frac{7}{20} \\
&= \frac{|A \cap B|}{|B|} && [|B| = 20 \text{ people like donuts}, |A \cap B| = 7 \text{ people like both}] \\
&= \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} && [\text{divide top and bottom by } |\Omega|\text{, which is equivalent}] \\
&= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} && [\text{if we have equally likely outcomes}]
\end{aligned}
$$

This intuition (which worked only in the special case equally likely outcomes), leads us to the definition of conditional probability:

Definition 2.2.1: Conditional Probability

The **conditional probability** of event A given that event B happened is:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

An equivalent and useful formula we can derive (by multiplying both sides by the denominator, $\mathbb{P}(B)$, and switching the sides of the equation is:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$$

Let's consider an important question: does $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)$? No!

This is a common misconception we can show with some examples. In the above example with ice cream, we showed already $\mathbb{P}(A \mid B) = \frac{7}{20}$, but $\mathbb{P}(B \mid A) = \frac{7}{36}$, and these are not equal.

Consider another example where W is the event that you are wet and S is the event you are swimming. Then, the probability you are wet given you are swimming, $\mathbb{P}(W \mid S) = 1$, as if you are swimming you are certainly wet. But, the probability you are swimming given you are wet, $\mathbb{P}(S \mid W) \neq 1$, because there are numerous other reasons you could be wet that don't involve swimming (being in the rain, showering, etc.).

2.2.2 Bayes Theorem

This brings us to Bayes Theorem:

Theorem 2.2.1: Bayes Theorem

Let A, B be events with nonzero probability. Then,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

Note that in the above $\mathbb{P}(A)$ is called the **prior**, which is our belief without knowing anything about event B. $\mathbb{P}(A \mid B)$ is called the **posterior**, our belief after learning that event B occurred.

This theorem is important because it allows to "reverse the conditioning"! Notice that both $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$ appear in this equation on opposite sides. So if we know $\mathbb{P}(A)$ and $\mathbb{P}(B)$ and can more easily calculate one of $\mathbb{P}(A \mid B)$ or $\mathbb{P}(B \mid A)$, we can use **Bayes Theorem** to derive the other.

Proof of Bayes Theorem. Recall the (alternate) definition of conditional probability from above:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) \tag{2.2.1}$$

Swapping the roles of A and B we can also get that:

$$\mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) \tag{2.2.2}$$

But, because $A \cap B = B \cap A$ (since these are the outcomes in both events A and B, and the order of intersection does not matter), $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$, so (2.2.1) and (2.2.2) are equal and we have (by setting the right-hand sides equal):

$$\mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$$

We can divide both sides by $\mathbb{P}(B)$ and get Bayes Theorem:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

Wow, I wish I was alive back then and had this important (and easy to prove) theorem named after me! □

Example(s)

We'll investigate two slightly different questions whose answers don't seem that they should be different, but are. Suppose a family has two children (whom at birth, were each equally likely to be male or female). Let's say a telemarketer calls home and one of the two children picks up.

1. If the child who responded was male, and says "Let me get my *older* sibling", what is the probability that both children are male?
2. If the child who responded was male, and says "Let me get my *other* sibling", what is the probability that both children are male?

Solution There are four equally likely outcomes, MM, MF, FM, and FF (where M represents male and F represents female). Let A be the event both children are male.

1. In this part, we're given that the *younger* sibling is male. So we can rule out 2 of the 4 outcomes above and we're left with MF and MM. Out of these two, in one of these cases we get MM, and so our desired probability is $1/2$.
   More formally, let this event be B, which happens with probability $2/4$ (2 out of 4 equally likely outcomes). Then, $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{2/4} = \frac{1}{2}$, since $P(A \cap B)$ is the probability both children are male, which happens in 1 out of 4 equally likely scenarios. This is because the older sibling's sex is *independent* of the younger sibling's, so knowing the younger sibling is male doesn't change the probability of the older sibling being male (which is what we computed just now).

2. In this part, we're given that *at least one sibling* is male. That is, out of the 4 outcomes, we can only rule out the FF option. Out of the remaining options MM, MF, and FM, only one has both siblings being male. Hence, the probability desired is $1/3$. You can do a similar more formal argument like we did above!

See how a slight wording change changed the answer? $\square$

We'll see a disease testing example later, which requires the next section first. If you test positive for a disease, how concerned should you be? The result may surprise you!

2.2.3 Law of Total Probability

Let's say you sign up for a chemistry class, but are assigned to one of three teachers randomly. Furthermore, you know the probabilities you fail the class if you were to have each teacher (from historical results, or word-of-mouth from classmates who have taken the class). Can we combine this information to compute the overall probability that you fail chemistry (before you know which teacher you get)? Yes - using the law of total probability below! We first need to define what a partition is.

Definition 2.2.2: Partitions

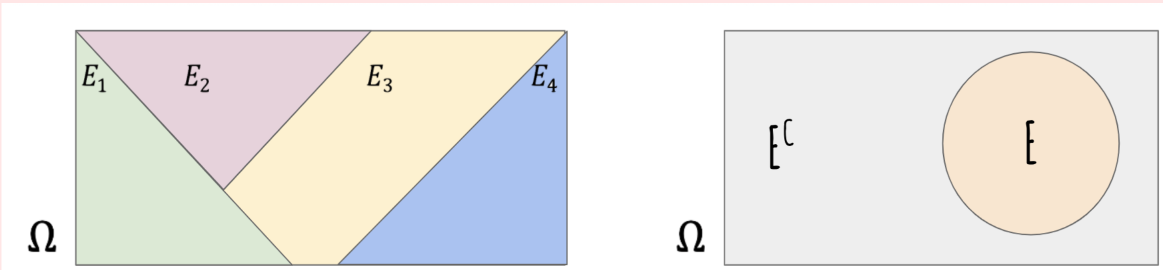
Non-empty events $E_1, \dots, E_n$ **partition** the sample space Ω if they are:

- **(Exhaustive)** $E_1 \cup E_2 \cup \dots \cup E_n = \bigcup_{i=1}^n E_i = \Omega$; that is, they cover the entire sample space.
- **(Pairwise Mutually Exclusive)** For all $i \neq j$, $E_i \cap E_j = \emptyset$; that is, none of them overlap.

Note that for any event E, E and E^C always form a partition of Ω.

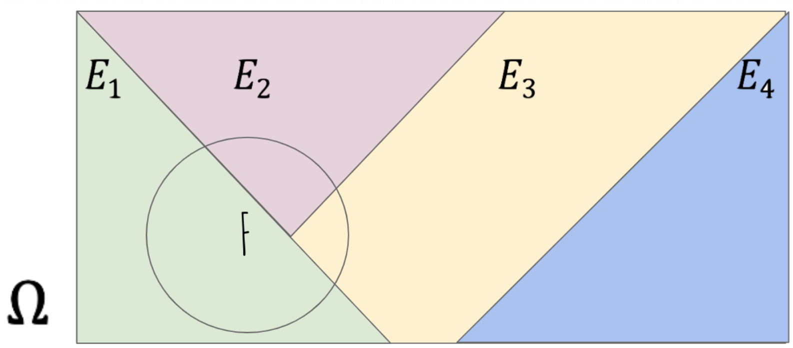
Example(s)

Two example partitions can be seen in the image below:



You can see that partition is a very appropriate word here! In the first image, the four events $E_1, \ldots, E_4$ don't overlap and cover the sample space. In the second image, the two events E, E^C do the same thing! This is useful when you know *exactly* one of a few things will happen. For example, for the chemistry example, there might be only three teachers, and you will be assigned to exactly one of them: at most one because you can't have two teachers (mutually exclusive), and at least one because there aren't other teachers possible (exhaustive).

Now, suppose we have some event F which intersects with various events that form a partition of Ω. This is illustrated by the picture below:



Notice that F is composed of its intersection with each of E_1, E_2, and E_3, and so we can split F up into smaller pieces. This means that we can write the following (green chunk $F \cap E_1$, plus pink chunk $F \cap E_2$ plus yellow chunk $F \cap E_3$):

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \mathbb{P}(F \cap E_2) + \mathbb{P}(F \cap E_3)$$

Note that F and E_4 do not intersect, so $F \cap E_4 = \emptyset$. For completion, we can include E_4 in the above equation, because $\mathbb{P}(F \cap E_4) = 0$. So, in all we have:

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \mathbb{P}(F \cap E_2) + \mathbb{P}(F \cap E_3) + \mathbb{P}(F \cap E_4)$$

This leads us to the law of total probability.

Theorem 2.2.2: Law of Total Probability (LTP)

If events $E_1, \ldots, E_n$ partition Ω, then for any event F

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \cdots + \mathbb{P}(F \cap E_n) = \sum_{i=1}^{n} \mathbb{P}(F \cap E_n)$$

Using the definition of conditional probability, $\mathbb{P}(F \cap E_i) = \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)$, we can replace each of the terms above and get the (typically) more useful formula:

$$\mathbb{P}(F) = \mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1) + \cdots + \mathbb{P}(F \mid E_n)\,\mathbb{P}(E_n) = \sum_{i=1}^{n} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)$$

That is, to compute the probability of an event F overall; suppose we have n disjoint cases $E_1, \ldots, E_n$ for which we can (easily) compute the probability of F in each of these cases ($P(F|E_i)$). Then, take the weighted average of these probabilities, using the probabilities $P(E_i)$ as weights (the probability of being in each case).

Example(s)

Let's consider an example in which we are trying to determine the probability that we fail chemistry. Let's call the event F failing, and consider the three events E_1 for getting the Mean Teacher, E_2 for getting the Nice Teacher, and E_3 for getting the Hard Teacher which partition the sample space. The following table gives the relevant probabilities:

	Mean Teacher E_1	Nice Teacher E_2	Hard Teacher E_3
Probability of Teaching You $\mathbb{P}(E_i)$	6/8	1/8	1/8
Probability of Failing You $\mathbb{P}(F \mid E_i)$	1	0	1/2

Solve for the probability of failing.

Solution Before doing anything, how are you liking your chances? There is a high probability (6/8) of getting the Mean Teacher, and she will certainly fail you. Therefore, you should be pretty sad.

Now let's do the computation. Notice that the first row sums to 1, as it must, since events E_1, E_2, E_3 partition the sample space (you have exactly one of the three teachers). Using the Law of Total Probability (LTP), we have the following:

$$\begin{aligned}\mathbb{P}(F) &= \sum_{i=1}^{3} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i) = \mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1) + \mathbb{P}(F \mid E_2)\,\mathbb{P}(E_2) + \mathbb{P}(F \mid E_3)\,\mathbb{P}(E_3) \\ &= 1 \cdot \frac{6}{8} + 0 \cdot \frac{1}{8} + \frac{1}{2} \cdot \frac{1}{8} = \frac{13}{16}\end{aligned}$$

Notice to get the probability of failing, what we did was: consider the probability of failing in each of the 3 cases, and take a weighted average of using the probability of each case. This is exactly what the law of total probability lets us do!

You might consider using the LTP when you know the probability of your desired event in □

Example(s)

Misfortune struck us and we ended up failing chemistry class. What is the probability that we had the Hard Teacher given that we failed?

Solution First, this probability should be low intuitively because if you failed, it was probably due to the Hard Teacher (because you are more likely to get them, AND because they have a high fail rate of 100%).

Start by writing out in a formula what you want to compute; in our case, it is $\mathbb{P}(E_3 \mid F)$ (getting the hard teacher **given** that we failed). We know $\mathbb{P}(F \mid E_3)$ and we want to solve for $\mathbb{P}(E_3 \mid F)$. This is a hint to use Bayes Theorem since we can reverse the conditioning! Using that with the numbers from the table and the previous question:

$$\begin{aligned} \mathbb{P}(E_3 \mid F) &= \frac{\mathbb{P}(F \mid E_3)\,\mathbb{P}(E_3)}{\mathbb{P}(F)} && \text{[bayes theorem]} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{8}}{\frac{13}{16}} \\ &= \frac{1}{13} \end{aligned}$$

□

2.2.4 Bayes Theorem with the Law of Total Probability

Oftentimes, the denominator in Bayes Theorem is hard, so we must compute it using the LTP. Here, we just combine two powerful formulae: Bayes Theorem and the Law of Total Probability:

Theorem 2.2.3: Bayes Theorem with the Law of Total Probability

Let events $E_1, \ldots, E_n$ partition the sample space Ω, and let F be another event. Then:

$$\begin{aligned} \mathbb{P}(E_1 \mid F) &= \frac{\mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1)}{\mathbb{P}(F)} && \text{[by bayes theorem]} \\ &= \frac{\mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1)}{\sum_{i=1}^{n} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)} && \text{[by the law of total probability]} \end{aligned}$$

In particular, in the case of a simple partition of Ω into E and E^C, if E is an event with nonzero probability, then:

$$\begin{aligned} \mathbb{P}(E \mid F) &= \frac{\mathbb{P}(F \mid E)\,\mathbb{P}(E)}{\mathbb{P}(F)} && \text{[by bayes theorem]} \\ &= \frac{\mathbb{P}(F \mid E)\,\mathbb{P}(E)}{\mathbb{P}(F \mid E)\,\mathbb{P}(E) + \mathbb{P}(F \mid E^C)\,\mathbb{P}(E^C)} && \text{[by the law of total probability]} \end{aligned}$$

2.2.5 Exercises

1. Suppose the llama flu disease has become increasingly common, and now 0.1% of the population has it (1 in 1000 people). Suppose there is a test for it which is 98% accurate (e.g., 2% of the time it will give the wrong answer). Given that you tested positive, what is the probability you have the disease? Before any computation, think about what you think the answer might be.

   Solution: Let L be the event you have the llama flu, and T be the event you test positive (T^C is the event you test negative). You are asked for $\mathbb{P}(L \mid T)$. We do know $\mathbb{P}(T \mid L) = 0.98$ because if you have the llama flu, the probably you test positive is 98%. This gives us the hint to use Bayes Theorem!

   We get that
$$\mathbb{P}(L \mid T) = \frac{\mathbb{P}(T \mid L)\,\mathbb{P}(L)}{\mathbb{P}(T)}$$

   We are given $\mathbb{P}(T \mid L) = 0.98$ and $\mathbb{P}(L) = 0.001$, but how can we get $\mathbb{P}(T)$, the probability of testing positive? Well that depends on whether you have the disease or not. When you have two or more cases (L and L^C), that's a hint to use the LTP! So we can write
$$\mathbb{P}(T) = \mathbb{P}(T \mid L)\,\mathbb{P}(L) + \mathbb{P}\left(T \mid L^C\right)\mathbb{P}\left(L^C\right)$$

   Again, interpret this as a weighted average of the probability of testing positive whether you had llama flu $\mathbb{P}(T \mid L)$ or not $\mathbb{P}\left(T \mid L^C\right)$, weighting by the probability you are in each of these cases $\mathbb{P}(L)$ and $\mathbb{P}\left(L^C\right)$. We know $\mathbb{P}\left(L^C\right) = 0.999$ since these $\mathbb{P}\left(L^C\right) = 1 - \mathbb{P}(L)$ (axiom of probability). But what about $\mathbb{P}\left(T \mid L^C\right)$? This is the probability of testing positive given that you don't have llama flu, which is 0.02 or 2% (due to the 98% accuracy). Putting this all together, we get:
$$\begin{aligned}
\mathbb{P}(L \mid T) &= \frac{\mathbb{P}(T \mid L)\,\mathbb{P}(L)}{\mathbb{P}(T)} && \text{[bayes theorem]}\\
&= \frac{\mathbb{P}(T \mid L)\,\mathbb{P}(L)}{\mathbb{P}(T \mid L)\,\mathbb{P}(L) + \mathbb{P}(T \mid L^C)\,\mathbb{P}(L^C)} && \text{[LTP]}\\
&= \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.02 \cdot 0.999}\\
&\approx 0.046756
\end{aligned}$$

   Not even a 5% chance we have the disease, what a relief! But wait, how can that be? The test is so accurate, and it said you were positive? This is because the prior probability of having the disease $\mathbb{P}(L)$ was so low at 0.1% (actually this is pretty high for a disease rate). If you think about it, the posterior probability we computed $\mathbb{P}(L \mid T)$ is $47\times$ larger than the prior probability $\mathbb{P}(L)$ ($\mathbb{P}(L \mid T) / \mathbb{P}(L) \approx 0.047/0.001 = 47$), so the test did make it a lot more likely we had the disease after all!

2. Suppose we have four fair die: one with three sides, one with four sides, one with five sides, and one with six sides (The numbering of an n-sided die is $1, 2, ..., n$). We pick one of the four die, each with equal probability, and roll the same die three times. We get all 4's. What is the probability we chose the 5-sided die to begin with?

   Solution: Let D_i be the event we rolled the i-sided die, for $i = 3, 4, 5, 6$. Notice that these

D_3, D_4, D_5, D_6 partition the sample space.

$$
\begin{aligned}
P(D_5|444) &= \frac{P(444|D_5)P(D_5)}{P(444)} && \text{[by bayes theorem]} \\
&= \frac{P(444|D_5)P(D_5)}{P(444|D_3)P(D_3) + P(444|D_4)P(D_4) + P(444|D_5)P(D_5) + P(444|D_6)P(D_6)} && \text{[by ltp]} \\
&= \frac{\frac{1}{5^3}\cdot\frac{1}{4}}{\frac{0}{3^3}\cdot\frac{1}{4} + \frac{1}{4^3}\cdot\frac{1}{4} + \frac{1}{5^3}\cdot\frac{1}{4} + \frac{1}{6^3}\cdot\frac{1}{4}} \\
&= \frac{1/125}{1/64 + 1/125 + 1/216} \\
&= \frac{1728}{6103} \approx 0.2831
\end{aligned}
$$

Note that we compute $P(444|D_i)$ by noting there's only one outcome where we get $(4, 4, 4)$ out of the i^3 equally likely outcomes. This is true except when $i = 3$, where it's not possible to roll all 4's.