Chapter 2. Discrete Probability

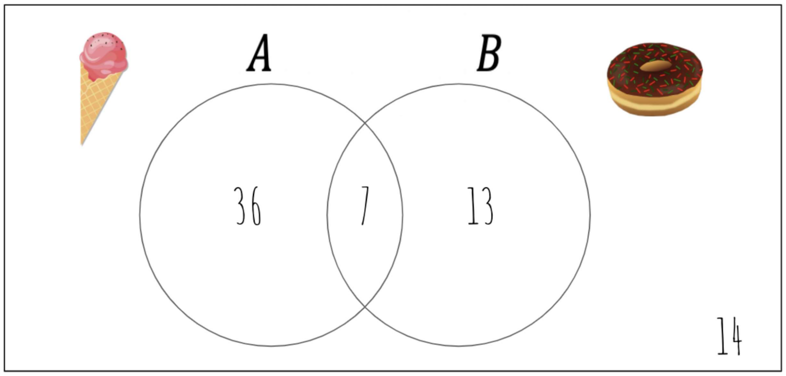
2.2: Conditional Probability

Slides (Google Drive) 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 $\mathbb{P}(D \mid T^C)$ 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 $\mathbb{P}(D \mid T^C) < \mathbb{P}(D)$ 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 2.2.8: 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:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) \tag{2.2.6}$$

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

$$\mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) \tag{2.2.7}$$

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? □

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

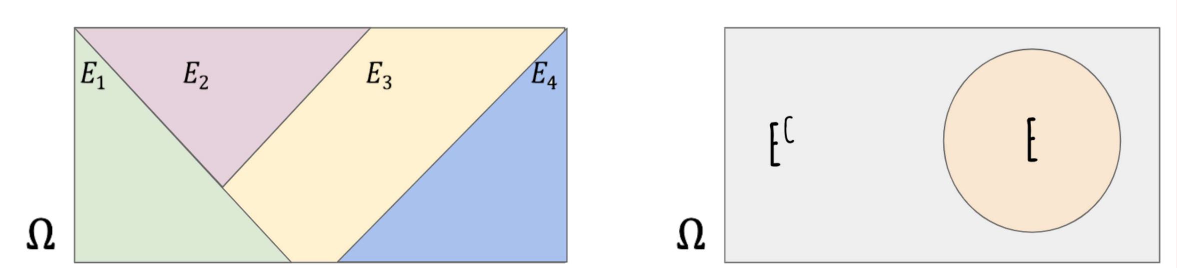
Non-empty events $E_1, \dots, E_n$ **partition** the sample space Ω if they are:

- **(Exhaustive)** $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^{n} E_i = \Omega$; 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.

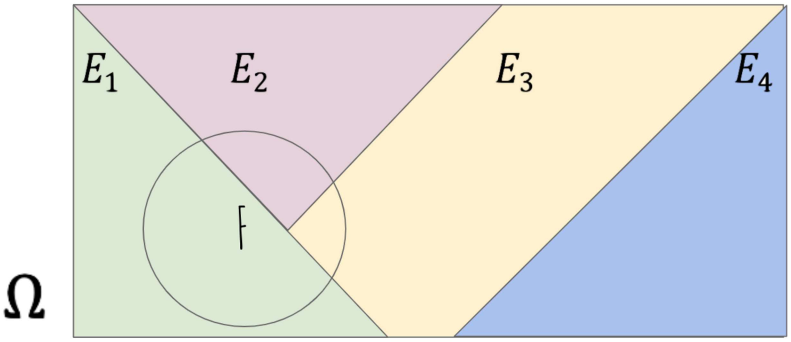
Example(s)

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 $E_1, \dots, E_4$ 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 2.2.9: 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! □

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 2.2.10: 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× 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.

Chapter 2. Discrete Probability

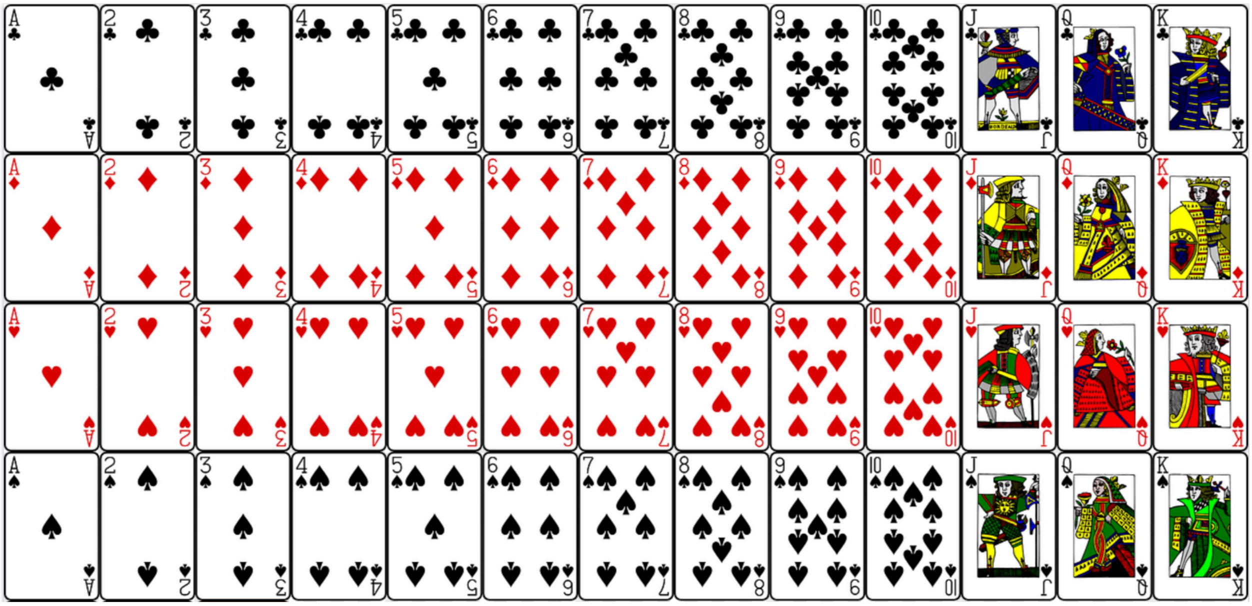
2.3: Independence

Slides (Google Drive) Video (YouTube)

2.3.1 Chain Rule

We learned several tools already to compute probabilities (equally likely outcomes, Bayes Theorem, LTP). Now, we will learn how to handle the probability of several events occurring simultaneously: that is, $\mathbb{P}(A \cap B \cap C \cap D)$ for example. To compute the probability that at least one of several events happens: $\mathbb{P}(A \cup B \cup C \cup D)$, you would use inclusion-exclusion! We'll see an example which builds intuition first.

Consider a standard 52 card deck. This has four suits (clubs, spades, hearts, and diamonds). Each of the four suits has 13 cards of different rank (A, 2, 3, 4, 5, 6, 7, 8, 9, 10 J, Q, K).



Now, suppose that we shuffle this deck and draw the top three cards. Let's define:

1. A to be the event that we get the Ace of spades as our **first** card.
2. B to be the event that we get the 10 of clubs as our **second** card.
3. C to be the event that we get the 4 of diamonds as our **third** card.

What is the probability that all three of these events happen? We can write this as $\mathbb{P}(A, B, C)$ (sometimes we use commas as an alternative to using the intersection symbol, so this is equivalent to $\mathbb{P}(A \cap B \cap C)$). Note that this is equivalent to $\mathbb{P}(C, B, A)$ or $\mathbb{P}(B, C, A)$ since order of intersection does not matter.

Intuitively, you might say that this probability is $\frac{1}{52} \cdot \frac{1}{51} \cdot \frac{1}{50}$, and you would be correct.

1. The first factor comes from the fact that there are 52 cards that could be drawn, and only one ace of spades. That is, we computed $\mathbb{P}(A)$.

2. The second factor comes from the fact that there are 51 cards after we draw the first card and only one 10 of clubs. That is, we computed $\mathbb{P}(B \mid A)$.
3. The final factor comes from the fact that there are 50 cards left after we draw the first two and only one 4 of diamonds. That is, we computed $\mathbb{P}(C \mid A, B)$.

To summarize, we said that

$$\mathbb{P}(A, B, C) = \mathbb{P}(A) \cdot \mathbb{P}(B \mid A) \cdot \mathbb{P}(C \mid A, B) = \frac{1}{52} \cdot \frac{1}{51} \cdot \frac{1}{50}$$

This brings us to the chain rule:

Theorem 2.3.11: Chain Rule

Let $A_1, \ldots, A_n$ be events with nonzero probabilities. Then:

$$\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 A_2) \cdots \mathbb{P}(A_n \mid A_1, \ldots, A_{n-1})$$

In the case of two events, A, B (this is just the alternate form of the definition of conditional probability from 2.2):

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

An easy way to remember this, is if we want to observe n events, we can observe one event at a time, and condition on those that we've done thus far. And most importantly, since the order of intersection **doesn't matter**, you can actually decompose this into any of $n!$ orderings. Make sure you "do" one event at a time, conditioning on the intersection of ALL past events like we did above.

Proof of Chain Rule. Remember that the definition of conditional probability says $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)$. We'll use this repeatedly to break down our $\mathbb{P}(A_1, \ldots, A_n)$. Sometimes it is easier to use commas, and sometimes it is easier to use the intersection sign $\cap$: for this proof, we'll use the intersection sign. We'll prove this for four events, and you'll see how it can be easily extended to any number of events!

$$\begin{aligned}
\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) &= \mathbb{P}((A_1 \cap A_2 \cap A_3) \cap A_4) && [\text{treat } A_1 \cap A_2 \cap A_3 \text{ as one event}] \\
&= \mathbb{P}(A_1 \cap A_2 \cap A_3)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)] \\
&= \mathbb{P}((A_1 \cap A_2) \cap A_3)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) && [\text{treat } A_1 \cap A_2 \text{ as one event}] \\
&= \mathbb{P}(A_1 \cap A_2)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_3 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)] \\
&= \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_3 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)]
\end{aligned}$$

Note how we keep "chaining" and applying the definition of conditional probability repeatedly!

□

Example(s)

Consider the 3-stage process. We roll a 6-sided die (numbered 1-6), call the outcome X. Then, we roll a X-sided die (numbered 1-X), call the outcome Y. Finally, we roll a Y-sided die (numbered 1-Y), call the outcome Z. What is $P(Z = 5)$?

Solution There are only three things that could have happened for the triplet (X, Y, Z) so that Z takes on the value 5: $\{(6,6,5),(6,5,5),(5,5,5)\}$. So

$$\begin{aligned}\mathbb{P}(Z=5) &= \mathbb{P}(X=6, Y=6, Z=5) + \mathbb{P}(X=6, Y=5, Z=5) + \mathbb{P}(X=5, Y=5, Z=5) && \text{[cases]} \\ &= \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} + \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{5} + \frac{1}{6}\cdot\frac{1}{5}\cdot\frac{1}{5} && \text{[chain rule 3x]}\end{aligned}$$

How did we use the chain rule? Let's see for example the last term:

$$\mathbb{P}(X=5, Y=5, Z=5) = P(X=5)P(Y=5 \mid X=5)P(Z=5 \mid X=5, Y=5)$$

$P(X=5) = \frac{1}{6}$ because we rolled a 6-sided die.
$P(Y=5 \mid X=5) = \frac{1}{5}$ since we rolled a $X=5$-sided die.
Finally, $P(Z=5 \mid X=5, Y=5) = P(Z=5 \mid Y=5) = \frac{1}{5}$ since we rolled a $Y=5$-sided die. Note we didn't need to know $X=5$ once we knew $Y=5$!

□

2.3.2 Independence

Let's say we flip a fair coin 3 times *independently* (whatever that means) - what is the probability of getting all heads? You may be inclined to say $(1/2)^3 = 1/8$ because the probability of getting heads each time is just $1/2$. However, we haven't learned such a rule to compute the joint probability $\mathbb{P}(H_1 \cap H_2 \cap H_3)$ except the chain rule.

Using only what we've learned, we could consider equally likely outcomes. There are $2^3 = 8$ possible outcomes when flipping a coin three times (by product rule), and only one of those (HHH) makes up the event we care about: $H_1 \cap H_2 \cap H_3$. Since the outcomes are equally likely,

$$\mathbb{P}(H_1 \cap H_2 \cap H_3) = \frac{|\, H_1 \cap H_2 \cap H_3 \,|}{|\, \Omega \,|} = \frac{|\, \{HHH\} \,|}{2^3} = \frac{1}{8}$$

We'd love a rule to say $\mathbb{P}(H_1 \cap H_2 \cap H_3) = \mathbb{P}(H_1) \cdot \mathbb{P}(H_2) \cdot \mathbb{P}(H_3) = 1/2 \cdot 1/2 \cdot 1/2 = 1/8$ - and it turns out this is true when the events are independent!

But first, let's consider the smaller case: does $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$ in general? No! How do we know this though? Well recall that by the chain rule, we know that:

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

So, unless $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ the equality does not hold. However, when this equality does hold, it is a special case, which brings us to independence.

Definition 2.3.1: Independence

Events A and B are **independent** if any of the following equivalent statements hold:
1. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
2. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
3. $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$

Intuitively what it means for $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ is that: given that we know B happened, the probability of observing A is the same as if we didn't know anything. So, event B has no influence on event A. The last statement

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

is the most often applied to problems where we are allowed to assume independence.

What about independence of more than just two events? We call this concept "mutual independence" (but most of the time we don't even say the word "mutual"). You might think that for events A_1, A_2, A_3, A_4 to be (mutually) independent, by extension of the definition of two events, we would just need

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4)$$

But it turns out, we need this property to hold for *any* subset of the 4 events. For example, the following must be true (in addition to others):

$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_3)$$

$$\mathbb{P}(A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4)$$

For all 2^n subsets of the 4 events ($2^4 = 16$ in our case), the probability of the intersection must simply be the product of the individual probabilities.

As you can see, it would be quite annoying to check even if three events were (mutually) independent. Luckily, most of the time we are told to assume that several events are (mutually) independent and we get all of those statements to be true for free. We are rarely asked to demonstrate/prove mutual independence.

Definition 2.3.2: Mutual Independence

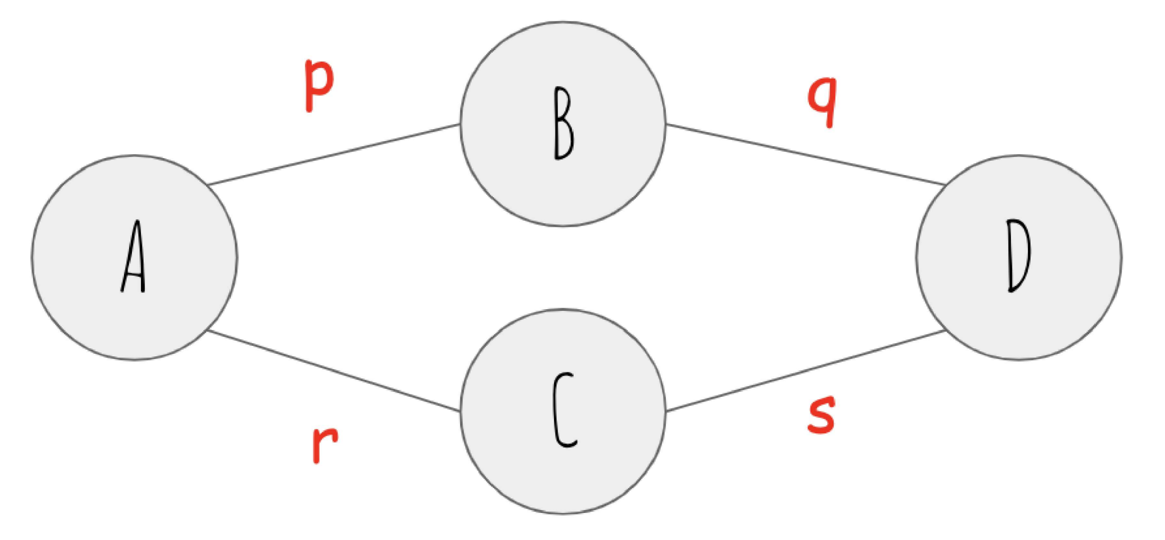
We say n events $A_1, A_2, \ldots, A_n$ are **(mutually) independent** if, for *any* subset $I \subseteq [n] = \{1, 2, \ldots, n\}$, we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$$

This is very similar to the last formula $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$ in the definition of independence for two events, just extended to multiple events. It must hold for any subset of the n events, and so this equation is actually saying 2^n equations are true!

Example(s)

Suppose we have the following network, in which circles represents a node in the network (A, B, C, and D) and the links have the probabilities p, q, r and s of successfully working. That is, for example, the probability of successful communication from A to B is p. Each link is independent of the others though.



Now, let's consider the question, what is the probability that A and D can successfully communicate?

Solution There are two ways in which it can communicate: (1) in the top path via B or (2) in the bottom path via C. Let's define the event *top* to be successful communication in the top path and the event *bottom* to be successful communication in the bottom path. Let's first consider the probabilities of each of these being successful communication. For the top to be a valid path, *both* links AB and BD must work.

$$\begin{aligned}\mathbb{P}(\text{top}) &= \mathbb{P}(AB \cap BD) \\ &= \mathbb{P}(AB)\,\mathbb{P}(BD) && \text{[by independence]} \\ &= pq\end{aligned}$$

Similarly:

$$\begin{aligned}\mathbb{P}(\text{bottom}) &= \mathbb{P}(AC \cap CD) \\ &= \mathbb{P}(AC)\,\mathbb{P}(CD) && \text{[by independence]} \\ &= rs\end{aligned}$$

So, to calculate the probability of successful communication between A and D, we can take the union of *top* and *bottom* (we just need at least one of the two to work), and so we have:

$$\begin{aligned}\mathbb{P}(\text{top} \cup \text{bottom}) &= \mathbb{P}(\text{top}) + \mathbb{P}(\text{bottom}) - \mathbb{P}(\text{top} \cap \text{bottom}) && \text{[by inclusion-exclusion]} \\ &= \mathbb{P}(\text{top}) + \mathbb{P}(\text{bottom}) - \mathbb{P}(\text{top})\,\mathbb{P}(\text{bottom}) && \text{[by independence]} \\ &= pq + rs - pqrs\end{aligned}$$

□

2.3.3 Conditional Independence

In the example above for the chain rule, we made this step:

$$\mathbb{P}(Z = 5 \mid X = 5, Y = 5) = \mathbb{P}(Z = 5 \mid Y = 5)$$

This is actually another form of independence, called conditional independence! That is, *given* that $Y = 5$, the events $X = 5$ and $Z = 5$ are independent (the above equation looks exactly like $\mathbb{P}(Z = 5 \mid X = 5) = \mathbb{P}(Z = 5)$ except with extra conditioning on $Y = 5$ on both sides.

Definition 2.3.3: Conditional Independence

Events A and B are **conditionally independent given an event** C if any of the following equivalent statements hold:

1. $\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$
2. $\mathbb{P}(B \mid A, C) = \mathbb{P}(B \mid C)$
3. $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$

Recall the definition of A and B being (unconditionally) independent below:

1. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
2. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
3. $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$

Notice that this is very similar to the definition of independence. There is no difference, except we have just added in conditioning on C to every probability.

Example(s)

Suppose there is a coin C_1 with $\mathbb{P}(head) = 0.3$ and a coin C_2 with $\mathbb{P}(head) = 0.9$. We pick one randomly with equal probability and will flip that coin 3 times *independently*. What is the probability we get all heads?

Solution Let us call HHH the event of getting three heads, C_1 the event of picking the first coin, and C_2 the event of getting the second coin. Then we have the following:

$$\begin{aligned}
\mathbb{P}(HHH) &= \mathbb{P}(HHH \mid C_1)\,\mathbb{P}(C_1) + \mathbb{P}(HHH \mid C_2)\,\mathbb{P}(C_2) && \text{[by the law of total probability]} \\
&= (\mathbb{P}(H \mid C_1))^3\mathbb{P}(C_1) + (\mathbb{P}(H \mid C_2))^3\mathbb{P}(C_2) && \text{[by conditional independence]} \\
&= (0.3)^3\frac{1}{2} + (0.9)^3\frac{1}{2} = 0.378
\end{aligned}$$

It is important to note that getting heads on the first and second flip are NOT independent. The probability of heads on the second, given that we got heads on the first flip, is much higher since we are more likely to have chosen coin C_2. However, *given which coin we are flipping*, the flips are conditionally independent. Hence, we can write $\mathbb{P}(HHH \mid C_1) = \mathbb{P}(H \mid C_1)^3$. □

2.3.4 Exercises

1. Corrupted by their power, the judges running the popular game show America's Next Top Mathematician have been taking bribes from many of the contestants. During each of two episodes, a given contestant is either allowed to stay on the show or is kicked off. If the contestant has been bribing the judges, she will be allowed to stay with probability 1. If the contestant has not been bribing the judges, she will be allowed to stay with probability 1/3, independent of what happens in earlier episodes. Suppose that 1/4 of the contestants have been bribing the judges. The same contestants bribe the judges in both rounds.

   (a) If you pick a random contestant, what is the probability that she is allowed to stay during the first episode?

   (b) If you pick a random contestant, what is the probability that she is allowed to stay during both episodes?

(c) If you pick a random contestant who was allowed to stay during the first episode, what is the probability that she gets kicked off during the second episode?

(d) If you pick a random contestant who was allowed to stay during the first episode, what is the probability that she was bribing the judge?

Solution:

(a) Let S_i be the event a contestant stays in the i^{th} episode, and B be the event a contestant is bribing the judges. Then, by the law of total probability,

$$\mathbb{P}(S_1) = \mathbb{P}(S_1 \mid B)\mathbb{P}(B) + \mathbb{P}\left(S_1 \mid B^C\right)\mathbb{P}\left(B^C\right) = 1 \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{2}$$

(b) Again by the law of total probability,

$$\begin{aligned}
\mathbb{P}(S_1 \cap S_2) &= \mathbb{P}(S_1 \cap S_2 \mid B)\mathbb{P}(B) + \mathbb{P}\left(S_1 \cap S_2 \mid B^C\right)\mathbb{P}\left(B^C\right) && \text{[LTP]} \\
&= \mathbb{P}(S_1 \mid B)\mathbb{P}(S_S \mid B)\mathbb{P}(B) + \mathbb{P}\left(S_1 \mid B^C\right)\mathbb{P}\left(S_2 \mid B^C\right)\mathbb{P}\left(B^C\right) && \text{[conditional independence]} \\
&= 1 \cdot 1 \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} \\
&= \frac{1}{3}
\end{aligned}$$

Again, it's important to note that staying on the first and second episode are NOT independent. If we know she stayed on the first episode, then it is more likely she stays on the second (since she's more likely to be bribing the judges). However, conditioned on whether or not we are bribing the judges, S_1 and S_2 are independent.

(c)

$$\mathbb{P}\left(S_2^C \mid S_1\right) = \frac{\mathbb{P}\left(S_1 \cap S_2^C\right)}{\mathbb{P}(S_1)}$$

The denominator is our answer to (a), and the numerator can be computed in the same way as (b).

(d) By Bayes Theorem,

$$\mathbb{P}(B \mid S_1) = \frac{\mathbb{P}(S_1 \mid B)\mathbb{P}(B)}{\mathbb{P}(S_1)}$$

We computed all these quantities in part (a).

2. A parallel system functions whenever at least one of its components works. Consider a parallel system of n components and suppose that each component works with probability p independently

(a) What is the probability the system is functioning?

(b) If the system is functioning, what is the probability that component 1 is working?

(c) If the system is functioning and component 2 is working, what is the probability that component 1 is working?

Solution:

(a) Let C_i be the event component i is functioning, for $i = 1, \ldots, n$. Let F be the event the system functions. Then,

$$\begin{aligned}
\mathbb{P}(F) &= 1 - \mathbb{P}\left(F^C\right) \\
&= 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} C_i^C\right) && \text{[def of parallel system]} \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}\left(C_i^C\right) && \text{[independence]} \\
&= 1 - (1-p)^n && \text{[prob any fails is } 1-p\text{]}
\end{aligned}$$

(b) By Bayes Theorem, and since $\mathbb{P}(F \mid C_1) = 1$ (system is guaranteed to function if C_1 is working),

$$P(C_1 \mid F) = \frac{\mathbb{P}(F \mid C_1)\,\mathbb{P}(C_1)}{\mathbb{P}(F)} = \frac{1 \cdot p}{1-(1-p)^n}$$

(c)

$$\begin{aligned}
\mathbb{P}(C_1 \mid C_2, F) &= \mathbb{P}(C_1 \mid C_2) && \text{[if given } C_2\text{, already know } F \text{ is true]} \\
&= \mathbb{P}(C_1) && [C_1, C_2 \text{ independent}] \\
&= p
\end{aligned}$$

Application Time!!

Now you've learned enough theory to discover the Naive Bayes classifier covered in section 9.3. You are highly encouraged to read that section before moving on!