1 Conditional Probability: Discrete Case

Let A be an event. Recall that $\mathbb{P}(A)$ is the probability that A occurs. Now, suppose that we're given some additional information. Then, this information may not completely determine whether A has occurred, but it might give us some valuable partial information.

(Example.) Suppose that a magician rolls a fair die. Let A be the event that we roll a 6. Then, we know that

$$\mathbb{P}(A) = \frac{1}{6}.$$

Now, let's suppose that the magician tells you that the result is an even number (either 2, 4, or 6), but does not yet reveal the full result of the roll to you. How does the probability of A change? It's certainly not $\frac{1}{6}$ since we know that it has to be an even number. Intuitively, the answer is $\frac{1}{3}$; this is particularly because the die originally was uniform, so there is no reason why – when the magician got an even number – the even numbers have a heavier weight.

1.1 What is Conditional Probaiblity?

Conditional probability is the study of probability under the presence of partial information.

Definition 1.1: Conditional Probability

Let A and B be two events such that

- A is the event of interest.
- B is the event that encodes the partial information that we have.

Suppose that $\mathbb{P}(B) > 0$. Then, the **conditional probability** of A, given that B has occurred, is denoted by $\mathbb{P}(A|B)$.

Remarks:

- We require $\mathbb{P}(B) > 0$; if $\mathbb{P}(B) = 0$, then this would imply that B would never occur anyways.
- All of the given information is on the *right* of the bar.

In our example above, A would be the event of interest and B is the event coming from the magician telling us that the roll was an even number.

1.2 Finding Conditional Probability

How do we find $\mathbb{P}(A|B)$?

(Example, Continued.) In the previous example, it's clear that the probability of rolling a 6 should change from $\frac{1}{6}$ to $\frac{1}{3}$ once we found out that the roll is an even number. Additionally, the roll is uniformly random – now, we know that there is one of *three* possible numbers, so it should still remain uniformly random, but just on the sample space $\{2, 4, 6\}$ instead of the original sample space $\{1, 2, 3, 4, 5, 6\}$.

So, if A is the event that we rolled a 6, and B is the even that we rolled an even number, notice that:

- $\mathbb{P}(A) = \frac{1}{6}$: This is the probability that we roll a 6.
- $\mathbb{P}(B) = \frac{1}{2}$: This is the probability that we will any even number.
- $\mathbb{P}(A \cap B) = \frac{1}{6}$: This is the probability that both events occur. Notice that A is a subset of B; therefore, $A \cap B = A$.

Intuitively, we expect $\mathbb{P}(A|B) = \frac{1}{3}$, and indeed we note that

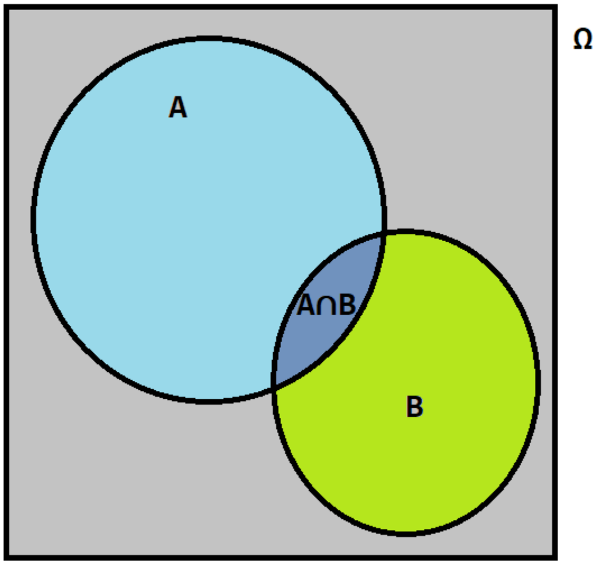
$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

In general, it is true that

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}.$$

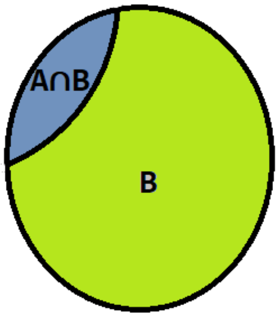
Why is this true?

(Informal Discussion.) Suppose that the sample space is Ω and we have two events $A, B \subset \Omega$. Then, consider the following Venn Diagram, where $\mathbb{P}(A)$ is represented by $\frac{\text{area}(A)}{\text{area}(\Omega)}$ and $\mathbb{P}(B)$ is represented by $\frac{\text{area}(B)}{\text{area}(\Omega)}$.



Now, the idea is that when we randomly throw a dart at the "dartboard" Ω, it will "land in" A with probability $\mathbb{P}(A)$ and similarly for B.

Now, suppose that we are told that the dart landed somewhere in B, but we don't know anything else beyond that. Then, since the dart was thrown randomly, it should be in some random position in B (i.e. nowhere in B should be more likely than anywhere else). Now, what is the probability that it landed in A, *given* that it landed somewhere in B?



The answer is to find out what is the probability that the dart landed in the intersection, that is, in the $A \cap B$ region. In this case, we can consider their *ratios* – in this case, we consider the area of $A \cap B$ against the area of B. Then, in effect, B becomes the new sample space (i.e. the “dartboard”), since we now know that the outcome of the experiment is in B. Therefore, we have

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

With this in mind, we can now give a formal definition.

Definition 1.2: Conditional Probability

Let A and B be two events. Suppose that $\mathbb{P}(B) > 0$. Then, the **conditional probability** of A, given that B has occurred, is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Recall that $\mathbb{P}(\omega)$ is a probability distribution on the sample space Ω *if* we have

- $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$, and
- $\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$.

Note that the function $\mathbb{P}(\omega|B)$ is *also* a probability distribution, but now the *sample space* is B. In particular:

1. $\mathbb{P}(\omega|B) = \frac{\mathbb{P}(\omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\omega)}{\mathbb{P}(B)} \geq 0$ for all $\omega \in B$. Note that this is true since $\mathbb{P}(B) > 0$ and $\mathbb{P}(\omega) \geq 0$ for all $\omega \in \Omega$.
2. $\sum_{\omega \in B} \mathbb{P}(\omega|B) = \frac{1}{\mathbb{P}(B)} \sum_{\omega \in B} \mathbb{P}(\omega) = \frac{1}{\mathbb{P}(B)} \mathbb{P}(B) = 1$.

Note that, by multiplying both sides of the conditional probability formula $P(A|B)$ by $P(B)$, we get the following formula:

Theorem 1.1: Probability Chain Rule

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B).$$

Remarks:

- The conditional probability formula only holds when $\mathbb{P}(B) > 0$.
- The probability chain rule $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B)$ holds even when $\mathbb{P}(B) = 0$.

(Example.) On sunny days, Vito goes for a walk with probability $\frac{4}{5}$. On rainy days, Vito goes for a walk with probability $\frac{1}{10}$. Suppose that, in San Diego, it rains on any given day with probability 3%. Find the probability that Vito goes for a walk today.

We let W be the event that Vito goes for a walk today; we want to find $\mathbb{P}(W)$. Let S be the event that it is sunny today.

We know that it rains with a 3% chance, so the change of it being sunny is 97%, or $\frac{97}{100}$. Thus, $\mathbb{P}(S) = \frac{97}{100}$. We also know that the probability that Vito goes for a walk, *given* that it is sunny[a], is $\mathbb{P}(W|S) = \frac{4}{5}$. Likewise, the probability that Vito goes for a walk, given that it is rainy (not a sunny day), is $\mathbb{P}(W|S^C) = \frac{1}{10}$.

Recall that, by the Law of Total Probability, we have

$$\mathbb{P}(W) = \mathbb{P}(W \cap S) + \mathbb{P}(W \cap S^C) \implies \mathbb{P}(W \cap S) = \mathbb{P}(W) - \mathbb{P}(W \cap S^C).$$

Additionally, we know that

$$\mathbb{P}(W \cap S^C) = \mathbb{P}(S^C)\mathbb{P}(W|S^C) = \frac{3}{100}\frac{1}{10} = \frac{3}{1000}.$$

Therefore, applying this formula to the probability chain rule and solving, we get the following work:

$$\begin{aligned}
\mathbb{P}(W \cap S) = \mathbb{P}(S)\mathbb{P}(W|S) & \quad \text{Probability Chain Rule} \\
\implies \mathbb{P}(W) - \mathbb{P}(W \cap S^C) = \mathbb{P}(S)\mathbb{P}(W|S) & \quad \text{Applying Law of Total Probability} \\
\implies \mathbb{P}(W) = \mathbb{P}(S)\mathbb{P}(W|S) + \mathbb{P}(W \cap S^C) & \quad \text{Add } \mathbb{P}(W \cap S^C) \text{ to Both Sides} \\
\implies \mathbb{P}(W) = \frac{97}{100}\frac{4}{5} + \frac{3}{1000} & \quad \text{Substitute Values} \\
\implies \mathbb{P}(W) = \frac{779}{1000} & \quad \text{Simplify}
\end{aligned}$$

[a] Note that we were not told that the probability of him going for a walk *and* it being sunny is $\frac{4}{5}$; all we're told is that *if* it is a sunny day, then he'll go for a walk with probability $\frac{4}{5}$.

1.3 Law of Total Probability, Conditional Version

Recall that, if $A \subset \Omega$ is an event, and if $B_1, \dots, B_n$ are partitions of Ω, then, the Law of Total Probability is given by

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i).$$

Using the probability chain rule, we have

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

This is the conditional version of the Law of Total Probability.

Theorem 1.2: Law of Total Probability, Conditional Version

Suppose that $A \subset \Omega$ is an event and that the events $B_1, \dots, B_n$ partition that sample space Ω. Then, we have that

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

1.4 Independent Events

Now that we have defined conditional probability, we can formally define what it means for events to be independent.

Definition 1.3: Independent Events

Two events A, B are **independent** if either

1. $\mathbb{P}(A) = 0$,
2. $\mathbb{P}(B) = 0$, or
3. $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$, in the case that both $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$.

Remark: With regards to (3), the occurrence of B does not affect the occurrence of A, and vice versa.

Now, by the probability chain rule, if two events A and B are independent, then

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

Indeed, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B)$. This introduces the next theorem:

Theorem 1.3

If A and B are independent, then

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

Now, independence is more complicated when more than two events are involved. In particular:

Definition 1.4

Events $A_1, \dots, A_n$ are **mutually independent** if, for any $A_{i_1}, \dots, A_{i_k}$, we have that

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_{i_k}\right) = \prod_{i=1}^{k} \mathbb{P}(A_{i_k}).$$

In particular, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i).$$

A weaker version onf independence is as follows:

Definition 1.5

Events $A_1, \dots, A_n$ are **pairwise independent** if each pair A_i, A_j (with $i \neq j$) are independent.

Note that if $n > 2$, then mutually and pairwise independence are **not** equivalent.

1.5 Joint Probability Distributions

Suppose that $X_1, \dots, X_n$ are discrete random variables with PMFs

$$p_{X_i} = \mathbb{P}(X_i = x).$$

Then, the joint PMF of the RV $\mathbf{X} = (X_1, \dots, X_n)$ is the function

$$p_{\mathbf{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

for all $x_1, \ldots, x_n$. Here, we can think of this as a function from a sample space to $\mathbb{R}^n$ (higher dimensions).

Often, it can be quite difficult to find the joint PMFs. However, in the special case that $X_1, \ldots, X_n$ are **independent**, it is just the product of the individual PMFs; that is,

$$p_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

We say that $X_1, \ldots, X_n$ are **independent and identically distributed**[1] (IID for short) if they are mutually independent, and all have the same distribution. An example of this is a sequence $X_1, \ldots, X_n$ of n Bernoulli(p) trials. In this case, each trial has the simple distribution

$$\mathbb{P}(X_i = 1) = p$$

and

$$\mathbb{P}(X_i = 0) = 1 - p.$$

In general, though, in such a process, X_i can have any given distribution. In any case, if $X_1, \ldots, X_n$ are IIDs, then their joint PMF takes the simple form known as a **product distribution**. In particular, since the X_i's are IIDs, they have the same PMFs $p(x) = \mathbb{P}(X_i) = x$ no matter what i is. Hence, the joint PMF is just

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i).$$

For example, in a sequence of Bernoulli(p) trials,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p^k (1-p)^{n-k},$$

if exactly k of the $x_i = 1$ and the other $n - k$ of the $x_i = 0$.

[1] This is known as **independent trial process** in the textbook.