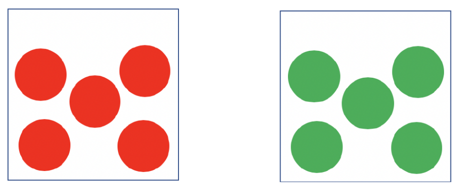
Lecture 4: Conditional probabilities, independence of events

In many situations, when we try to understand the probability of an event, we are often given some partial information. The presence of additional information will generally affect our judgement on the probability of an event.

We can illustrate this point by the following simple example.



Suppose that there are two urns, the first one containing 5 red balls and the second one containing 5 green balls. We select a ball randomly from the pool of the 10 balls. Without any information, we know that the probability of selecting a red ball is $\frac{1}{2}$. Now suppose we are given the extra information that the ball selected comes from the first urn. Under this new information, the probability of getting a red ball becomes 1, because the balls in the first urn are all red! This simple extreme example already indicates how the presence of additional information changes the judgement on probability.

To make the point mathematically precise, we are led to the study of conditional probabilities. The technique of conditioning is particularly important in modern probability theory, for instance in the study of Markov processes and martingales.

1 Two illustrative examples

Before defining conditional probabilities properly, let us look at another example.

Example 1.1. Toss two fair dice. The sample space is given by

$$\Omega = \{(1,1), (1,2) \cdots, (6,6)\}$$

which contains 36 elements. Consider the event that "the sum of the two dice is 8", namely,

$$A = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}.$$

Since this is a classical probability model, we see that

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{5}{36}.$$

Now let us assume that we are given the extra information that "the first die is a 3". Mathematically, this information is summarised by another event

$$H = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\}.$$

What should the probability of A be, *given this new information* (*i.e. given the occurrence of* H)? The point is that, under this new information, our sample space is changed to H since the only relevant outcomes are the ones where the first die is a 3. It is apparently that, all these outcomes should occur equally likely. In addition, there is precisely one outcome, namely $(3,5)$, among the six outcomes in H that triggers the event A. Therefore, we should expect that, the "conditional probability" of A given H is $\frac{1}{6}$. This is different from the unconditional probability that we have just calculated.

Example 1.2. The above example can be generalised to give a more illustrative discussion in the context of classical probability models. Let Ω be a finite set, and we consider the classical probability model over Ω. Let A, H be two events. We know that the "unconditional probability" of A is given by

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega}.$$

To define the "conditional probability" of A given H, just like Example 1.1 we are now restricting on the "new" sample space H. Among all the outcomes in H, the ones that trigger A are the ones in the event $A \cap H$. Therefore, under the classical probability model, it is natural to expect that, the "conditional probability" of A given H should be defined by

$$\mathbb{P}(A|H) = \frac{\#(A \cap H)}{\#H} = \frac{\#(A \cap H)/\#\Omega}{\#H/\#\Omega} = \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)}.$$

We can easily check that, this is exactly what is happening in Example 1.1, i.e.

$$\frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)} = \frac{1/36}{1/6} = \frac{1}{6}$$

in that example.

2 Definition of conditional probabilities

The previous discussion can go beyond the setting of classical probability models, leading us to the general definition of conditional probabilities. The general idea is the following.

Let A and H be two events. Suppose that we perform the random experiment for a total of n times (n is very large). Let n_1 be the number of times that H occurs, and let n_2 be the number of times that $A \cap H$ occurs. Apparently, $\frac{n_2}{n_1}$ gives the proportion of times that A occurs among those experiments where H occurs. Based on the heuristic principle that "relative frequencies should stabilise at the theoretical probability as $n \to \infty$", we should expect that:

$$\frac{n_2}{n_1} \text{ approximates the "theoretical conditional probability } \mathbb{P}(A|H)\text{''}$$

when $n \to \infty$. On the other hand, we know that

$$\frac{n_2}{n} \to \mathbb{P}(A \cap H), \ \frac{n_1}{n} \to \mathbb{P}(H) \quad \text{as } n \to \infty$$

due to the same principle. As a result, we have

$$\frac{n_2}{n_1} = \frac{n_2/n}{n_1/n} \to \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)} \quad \text{as } n \to \infty.$$

Therefore, it is reasonable to take the expression $\frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)}$ as the *definition* of the conditional probability of A given the occurrence of H. Of course this is only meaningful when $\mathbb{P}(H) > 0$.

Definition 2.1. Let A and H be two events, and assume that $\mathbb{P}(H) > 0$. The *conditional probability* of A given the occurrence of H, is defined by

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

Remark 2.1. Using the definition of $\mathbb{P}(A|H)$, we have

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

This is often known as the *multiplication theorem.*

Example 2.1. (i) Toss a die. Consider the events

$$A = \{2\}, \ H = \{2, 4, 6\}.$$

Then

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

(ii) Toss a pair of dice. Consider the events

$$A = \{(i, j) : |i - j| = 1\}, \ H = \{(i, j) : i + j = 7\}.$$

It is a simple matter of enumeration and counting to see that

$$\mathbb{P}(A) = \frac{5}{18}, \ \mathbb{P}(A|H) = \frac{1}{3}.$$

In both (i) and (ii) of the previous example, we see that $\mathbb{P}(A|H) > \mathbb{P}(A)$. In other words, the occurrence of H increases the chance that A happens. In this case, we say that there is a *positive relation* between A and H. Note that the roles of A and H is symmetric in this definition, since

$$\mathbb{P}(A|H) > \mathbb{P}(A) \iff \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)} > \mathbb{P}(A) \iff \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(A)} > \mathbb{P}(H) \iff \mathbb{P}(H|A) > \mathbb{P}(H).$$

Correspondingly, we say that there is a *negative relation* between A and H if $\mathbb{P}(A|H) < \mathbb{P}(A)$.

Let us look at a more practical example.

Example 2.2. You are going to interview a job. There are 7 candidates in total, and 3 positions will be offered. In this case, a generic outcome is given by three persons (those receiving the offer) out of the 7 candidates. The sample space Ω thus contains $\binom{7}{3} = 35$ outcomes in total.

Suppose that, among the 7 candidates, you and Ms X have a particular skill that the job needs. Therefore, it is likely that one of you will be employed, but not likely that both of you will get the offer. As a result, it is not reasonable to work with the classical probability model over Ω (i.e. assuming all outcomes occur equally likely). Instead, let us presume that, the underlying probability function (the a priori probabilities) is given in the following way.

(i) Each outcome corresponding to the event that "You and Ms X are both employed" has probability $\frac{1}{60}$. Simple counting shows that there are $\binom{5}{1} = 5$ outcomes in this event. So this event has probability $\frac{5}{60}$.
(ii) Each outcome corresponding to the event that "You get the job but Ms X does not" has probability $\frac{1}{24}$. There are $\binom{5}{2} = 10$ elements in this event, so the event has probability $\frac{10}{24}$.
(iii) Each outcome corresponding to the event that "You do not get the job but Ms X does" has probability $\frac{1}{24}$. There are $\binom{5}{2} = 10$ elements in this event, so the event has probability $\frac{10}{24}$.
(iv) Each outcome corresponding to the event that "neither you nor Ms X gets the job" has probability $\frac{1}{120}$. There are $\binom{5}{3} = 10$ elements in this event, so the event has probability $\frac{10}{120}$.

It is clear that, the above specifications define a legal probability function.What is the conditional probability that you are employed, given that Ms X is not employed?

Solution. Let us introduce the following events:

$$A : \text{You are employed}, \quad H : \text{Ms X is employed}.$$

The desired conditional probability is $\mathbb{P}(A|H^c)$. By the definition of the underlying probability function, we have

$$\mathbb{P}(A \cap H^c) = \frac{10}{24} = \frac{5}{12} \quad \text{(Case (ii))}$$

and

$$\begin{aligned}\mathbb{P}(H^c) &= \mathbb{P}(A \cap H^c) + \mathbb{P}(A^c \cap H^c) \quad \text{(Case (ii) + Case (iv))} \\ &= \frac{10}{24} + \frac{10}{120} = \frac{1}{2}.\end{aligned}$$

Therefore,

$$\mathbb{P}(A|H^c) = \frac{\mathbb{P}(A \cap H^c)}{\mathbb{P}(H^c)} = \frac{5/12}{1/2} = \frac{5}{6}.$$

We can also compute

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap H) + \mathbb{P}(A \cap H^c) \quad \text{(Case (i) + Case (ii))} \\ &= \frac{5}{60} + \frac{10}{24} = \frac{60}{120} = \frac{1}{2}.\end{aligned}$$

As a result, we see that given Ms X does not get the job, your chance is substantially increased. But this is not too surprising from the a priori assumptions.

3 Independence of events

We have seen the notions of positive / negative relations between two events. What if it is the case that $\mathbb{P}(A|H) = \mathbb{P}(A)$? This is saying that, the occurrence of H does not affect the probability of A. In other words, it is suggesting some sort of "independence" between the two events. Recall that, the property $\mathbb{P}(A|H) = \mathbb{P}(A)$ is algebraically equivalent to

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

Definition 3.1. Two events A, B are said to be *independent*, if

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

Two events are said to be *dependent* if they are not independent.

Example 3.1. Toss a die. Consider the following events:

$$A = \{2, 4, 6\},\ B = \{1, 2, 3\},\ C = \{1, 2, 3, 4\}.$$

We have

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

and

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

Therefore,

$$\mathbb{P}(A \cap B) \neq \mathbb{P}(A) \cdot \mathbb{P}(B),\ \mathbb{P}(A \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(C).$$

In other words, A, B are dependent, and A, C are independent.

Remark 3.1. So far we have seen two notions regarding some sort of "non-relatedness" between two events: independence and disjointness. These two notions are entirely different. On the one hand, to test disjointness one only needs to inspect the elements in the two events to see if there are common elements. This process does not involve any assigned probabilities. On the other hand, testing independence must involve the underlying probability function, which is apparent from the definition. One can still say something about their relation though. Suppose that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. If A, B are disjoint, then

$$0 = \mathbb{P}(\emptyset) = \mathbb{P}(A \cap B) < \mathbb{P}(A) \cdot \mathbb{P}(B),$$

since the right hand side is strictly positive. In other words, A, B must not be independent. More precisely, in this case A, B have a negative relation: $0 = \mathbb{P}(A|B) < \mathbb{P}(A)$. This is not surprising since the occurrence of B excludes the possibility that A happens (by disjointness).

Property. If A, B are independent, so are any of the following pairs:

$$\text{(i) } A^c \text{ and } B, \quad \text{(ii) } A \text{ and } B^c, \quad \text{(iii) } A^c \text{ and } B^c.$$

Proof. We only verify (i) and leave the other two as exercises. First of all, we have the decomposition

$$B = (A \cap B) \cup (A^c \cap B).$$

It follows that

$$\begin{aligned}
\mathbb{P}(A^c \cap B) &= \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \text{(by finite additivity)} \\
&= \mathbb{P}(B) - \mathbb{P}(A) \cdot \mathbb{P}(B) \quad \text{(by the independence of } A, B) \\
&= (1 - \mathbb{P}(A)) \cdot \mathbb{P}(B) \\
&= \mathbb{P}(A^c) \cdot \mathbb{P}(B) \quad \text{(by Property 3 of probability functions).}
\end{aligned}$$

Therefore, we conclude that A^c and B are independent. □

Independence of more than two events

The next natural question is to extend the notion of independence to the context of more than two events, say $A_1, A_2, \cdots, A_n$. This first naive attempt of such extension is to require *pairwise independence* among this events, namely, by requiring that

$$A_i, A_j \text{ are independent for any pair of } i \neq j.$$

Let us use a simple example to illustrate why this is *not* sufficient to suggest the "total independence" among these events.

Example 3.2. Toss two fair coins. Consider the following three events:

$$A : \text{first coin is Head}, \;\; B : \text{second coin is Head},$$
$$C : \text{exactly one of the two coins is Head}.$$

Simple calculation shows that

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2},$$

and

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{4}.$$

As a result, we see that these three events are pairwisely independent. However, they do share some sort of dependence as seen in the following way. Suppose that we know $A \cap B$ happens, i.e. both coins result in Head. This information immediately excludes the possibility of C. In other words,

$$\mathbb{P}(C|A \cap B) = 0 \neq \mathbb{P}(C).$$

This suggests dependence between $A \cap B$ and C, thus revealing certain dependence among the three events.

The above example tells us that, pairwise independent is not sufficient to capture "total independence". Indeed, we need to verify a lot more equations to establish independence.

Definition 3.2. Let $A_1, A_2, \cdots, A_n$ be n events. We say that $A_1, \cdots, A_n$ are *mutually independent*, if for any sub-collection of these events, namely for any sub-collection $\{j_1, \cdots, j_m\} \subseteq \{1, \cdots, n\}$, the following identity holds true:

$$\mathbb{P}(A_{j_1} \cap \cdots \cap A_{j_m}) = \mathbb{P}(A_{j_1}) \times \cdots \times \mathbb{P}(A_{j_m}).$$