Lecture 3: Properties of probability functions, the classical probability model

1 Continued from Lecture 2: further properties of probability functions

Property 3. For any event A, we have $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
Proof. Since $A \cup A^c = \Omega$, according to Axiom 2 and the finite additivity property (Property 2) we just shown, we have

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c).$$

Property 4. If $A \subseteq B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.
Proof. We have the decomposition $B = A \cup (A^c \cap B)$ as seen by:

$$\begin{aligned} A \cup (A^c \cap B) &= (A \cup A^c) \cap (A \cup B) \\ &= \Omega \cap B \quad (\text{since } A \subseteq B) \\ &= B. \end{aligned} \tag{1.1}$$

This decomposition is intuitively clear when one looks at a Venn diagram. Note that A and $A^c \cap B$ are disjoint. According to finite additivity, we have

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B). \tag{1.2}$$

In addition, from Axiom 1 we know that probabilities are always non-negative. In particular, the second term on the right hand side of (1.2) is non-negative, hence yielding $\mathbb{P}(B) \geqslant \mathbb{P}(A)$.
An important consequence of Property 4 is that probabilities are always taking values between 0 and 1, i.e. $0 \leqslant \mathbb{P}(A) \leqslant 1$ for every event A. This can be seen from the relation $\emptyset \subseteq A \subseteq \Omega$. It is interesting to point out that, this rather convincing fact is a deduced property instead of a presumed axiom.

The next property generalised finite additivity (for two events) to the case without assuming disjointness. This is known as the *Addition Theorem*, which is very useful in practice.

Property 5. For any two events A and B, we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{1.3}$$

Proof. We first consider the decomposition

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

Note that A and $B \cap A^c$ are disjoint. Therefore, we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B). \tag{1.4}$$

Next, we consider the decomposition

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

The events $A \cap B$ and $A^c \cap B$ are also disjoint. Therefore,

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B),$$

or equivalently,

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{1.5}$$

By substituting (1.5) into (1.4), we obtain the equation (1.3). The two decompositions we have considered here can also be heuristically seen by drawing a Venn diagram, and precisely checked by using the distributive laws (similar to the steps for reaching (1.1)).

The last property is concerned with the continuity of probability functions with respect to monotone sequences of events.

Property 6 [Continuity]. (i) Let $\{A_n : n \geqslant 1\}$ be an increasing sequence of events, namely, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, and define $A = \cup_{n=1}^{\infty} A_n$. Then

$$\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

(ii) Let $\{A_n : n \geqslant 1\}$ be an decreasing sequence of events, namely, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, and define $A = \cap_{n=1}^{\infty} A_n$. Then

$$\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

Proof. (i) We can write A as the union of mutually disjoint events:

$$A = A_1 \cup (A_2 \backslash A_1) \cup (A_3 \backslash A_2) \cup \cdots .$$

To simplify notation, let us denote $B_n = A_n \backslash A_{n-1}$. By Axiom 3, we have

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k).$$

But we also know that

$$A_n = B_1 \cup B_2 \cup \cdots \cup B_n,$$

and by finite additivity we have

$$\mathbb{P}(A_n) = \sum_{k=1}^{n} \mathbb{P}(B_k).$$

Therefore,

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(ii) This can be reduced to the case of Part (i) by taking complement ($\{A_n^c : n \geqslant 1\}$ is an increasing sequence of events whose union is A^c) and using Property 3.

2 The classical probability model

The simplest type of examples for probability functions is the so-called classical probability model.

Definition 2.1. A *classical probability model* refers to the following situation:

(i) The sample space Ω is a finite set, say having a total of N outcomes.
(ii) The probability function is defined in the way such that all outcomes in Ω occur equally likely. Namely,

$$\mathbb{P}(\{\omega\}) = \frac{1}{N} \quad \text{for every } \omega \in \Omega.$$

More generally, for any event $A \subseteq \Omega$, its probability is given by

$$\mathbb{P}(A) = \frac{\#A}{N},$$

where the notation $\#A$ means the number of elements in A.

3 Two enlightening examples

Many interesting examples in elementary probability theory fall in the category of the classical probability model. Let us look at two enlightening examples.

A Coin tossing example

We first consider the random experiment of tossing a coin. The sample space is $\Omega = \{H, T\}$. All possible events are given by

$$\emptyset, \{H\}, \{T\}, \Omega.$$

If $\mathbb{P}(\cdot)$ is a probability function, it has to satisfy

$$\mathbb{P}(\Omega) = 1,\ \mathbb{P}(\emptyset) = 0,$$

and

$$\mathbb{P}(\{T\}) = 1 - \mathbb{P}(\{H\}).$$

Therefore, to specify a legal probability function, it is enough to assign a value to $\mathbb{P}(\{H\})$.

If in the practical situation we are tossing a fair coin, we should assign $\frac{1}{2}$ to $\mathbb{P}(\{H\})$. This uniquely defines a probability function $\mathbb{P}$:

$$\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2},\ \mathbb{P}(\Omega) = 1.$$

The resulting model is a classical probability model.

On the other hand, if the coin is biased, say Heads occurs more often than Tails from empirical data, then we may assign some number $p \in (0.5, 1)$ to the probability of $\{H\}$. This defines a different probability function $\mathbb{Q}$:

$$\mathbb{Q}(\emptyset) = 0,\ \mathbb{Q}(\{H\}) = p,\ \mathbb{Q}(\{T\}) = 1 - p,\ \mathbb{Q}(\Omega) = 1,$$

which is also a legal probability function as it satisfies Kolmogorov's axioms. However, the resulting model is *not* a classical probability model.

Next, let us consider the experiment of toss a pair of fair coins. We want to compute the probability of "having a Head and a Tail". The situation is a bit subtler here, as there are two viewpoints we can take.

Viewpoint 1. We think of the two coins are ordered, say they are labeled by Coin 1 and Coin 2. Then the sample space is given by

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\},$$

where for a generic outcome $(\cdot, \cdot)$, the first component records the result of tossing Coin 1 and the second component records the result of tossing Coin 2. Since both coins are fair, we should consider the classical probability model over Ω. In other words, the probability function is defined in the way that the four outcomes occur equally likely:

$$\mathbb{P}(\{(H,H)\}) = \mathbb{P}(\{(H,T)\}) = \mathbb{P}(\{(T,H)\}) = \mathbb{P}(\{(T,T)\}) = \frac{1}{4}.$$

Under this model, the event "having a Head and a Tail" is given by

$$A = \{(H,T),(T,H)\},$$

and we have

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{2}{4} = \frac{1}{2}.$$

Viewpoint 2. In the 18th century, a famous mathematician D'Alembert had also considered this problem. His argument went as follows. Since we are tossing a pair of coins, there are three outcomes in total:

$$\Omega = \{H\&H, H\&T, T\&T\}.$$

Apparently, he was treating the two coins as identical and did not care about their order. This was also reasonable. Next, he claimed that, since the coins were fair, we should consider the corresponding classical probability model, namely defining the probability function such that

$$\mathbb{P}(\{H\&H\}) = \mathbb{P}(H\&T) = \mathbb{P}(T\&T) = \frac{1}{3}.$$

From this, it is apparent that the probability of "having a Head and a Tail" is $\frac{1}{3}$. D'Alembert's solution is apparently different from the first viewpoint, but both seem to be reasonable. What is happening here?

Unfortunately, this question cannot be answered from a pure mathematical perspective, and one needs to return to reality for empirical evidence. The reality is that, if one picks out two coins from his/her pocket and throw it for 500 times, it will be observed that the frequency of having a Head and a Tail is very close to $\frac{1}{2}$. This is clearly suggesting that Viewpoint 1 is more consistent with reality than Viewpoint 2. Therefore, Viewpoint 1 is considered a *practically correct* model.

However, we should remark that, Viewpoint 2 is not theoretically wrong! Indeed, from a mathematical perspective, both viewpoints are *theoretically correct*

because the underlying probability functions are both legal (i.e. satisfying Kolmogorov's axioms). The only issue of Viewpoint 2 is that it does not correspond to the reality we are considering. There is an easy way to fix this viewpoint without changing the sample space, so that we obtain the practically correct solution. Instead of considering the classical probability model, we should define

$$\mathbb{P}(\{H\&H\}) = \frac{1}{4},\ \mathbb{P}(\{H\&T\}) = \frac{1}{2},\ \mathbb{P}(\{T\&T\}) = \frac{1}{4}.$$

In this way, we get the same answer as Viewpoint 1 for the desired probability. Clearly, the underlying model here is *not* a classical probability model.

The birthday problem

In a group of n people, what is the probability that at least two of them have the same birthday? We may implicitly assume that, these people are independent and any day of the year is equally likely to be the birthday of a person.

To solve this problem, we can think of the n-people being labelled by $1, 2, \cdots, n$. The sample space can be written as

$$\Omega = \{(d_1, \cdots, d_n) : d_i = 1, 2, \cdots, 365 \text{ for each } 1 \leqslant i \leqslant n\}.$$

For a generic outcome $(d_1, \cdots, d_n) \in \Omega$, the i-th component d_i records the birthday of Person i. By the assumption, we should consider the classical probability model over Ω. Namely, each individual outcome occurs equally likely with probability

$$\mathbb{P}(\{(d_1, \cdots, d_n)\}) = \frac{1}{\#\Omega} = \frac{1}{365^n}.$$

Now let A be the event that "at least two people have the same birthday". Here the trick is that, it is much easier to compute $\mathbb{P}(A^c)$ rather than $\mathbb{P}(A)$. Indeed, we can easily count the number of elements in the event

$$A^c : \text{no two people have the same birthday.}$$

For Person 1, there are 365 possibilities. For Person 2, it leaves 364 possibilities for him/her to have a different birthday from Person 1, and similarly 363 possibilities for Person 3, 362 for Person 4 and so forth. Simple counting principle (see the *rule of product* in the supplementary notes "How do we count?") tells us that, the total number of elements in A^c is

$$\#A^c = \underbrace{365 \times 364 \times \cdots \times (365 - n + 1)}_{n \text{ terms}}.$$

Therefore, the desired probability is given by

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}. \tag{3.1}$$

The above expression is hard to evaluate in practice since 365^n is a very large number. There is a useful trick to numerically estimate this expression. We write

$$\begin{aligned}
\mathbb{P}(A^c) &= \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n} \\
&= \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365 \times 365 \times \cdots \times 365} \\
&= 1 \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - (n-1)}{365} \\
&= 1 \times \big(1 - \frac{1}{365}\big) \times \big(1 - \frac{2}{365}\big) \times \cdots \times \big(1 - \frac{n-1}{365}\big) \\
&= \exp\big(\sum_{k=1}^{n-1} \log\big(1 - \frac{k}{365}\big)\big),
\end{aligned}$$

where the last step follows from the identities

$$a = e^{\log a}, \ \log(a_1 a_2 \cdots a_m) = \sum_{k=1}^{m} \log a_k.$$

Next, since $\frac{k}{365}$ is small (if n is not too large), we can use the approximation

$$\log(1 - x) \approx -x \quad \text{when } x \text{ is small}$$

from calculus. This leads us to

$$\mathbb{P}(A^c) \approx \exp\big(-\frac{1}{365}\sum_{k=1}^{n-1} k\big) = \exp\big(-\frac{n(n-1)}{730}\big).$$

Therefore, we have

$$\mathbb{P}(A) \approx 1 - \exp\big(-\frac{n(n-1)}{730}\big), \tag{3.2}$$

which is much simpler than the original expression for calculation! This turns out to be a satisfactory approximation when n is not large.

By taking $n = 23$ either in (3.1) or (3.2), one finds that $\mathbb{P}(A) \approx 0.5$. In other words, in a group with only 23 people, the probability that "at least two people have the same birthday" is nearly 0.5 (Surprise?)!