Chapter 1

Probability Model

1.1 Experiments

Ideas about chance events and random behaviour arose out of thousands of years of game playing, long before any attempt was made to use mathematical reasoning about them. Board and dice games were well known in Egyptian times, and Augustus Caesar gambled with dice. Calculations of odds for gamblers were put on a proper theoretical basis by Fermat and Pascal in the early 17th century.

Definition 1.1. *An* **experiment** *is an activity or procedure that produces distinct, well-defined possibilities called* **outcomes**. *The set of all outcomes is called the* **sample space**, *and is denoted by* Ω.

The subsets of Ω *are called* **events**. *A single outcome,* ω, *when seen as a subset of* Ω, *as in* $\{\omega\}$, *is called a* **simple event**.

Given an outcome $\omega \in \Omega$ *we say that the event* $E \subset \Omega$ **occured** *if* $\omega \in E$.

Events, E_1, E_2 ... E_n, *that cannot occur at the same time are called* **mutually exclusive** *events, or* **pair-wise disjoint** *events. This means that* $E_i \cap E_j = \emptyset$ *where* $i \neq j$.

Example 1.2. *Some standard examples of experiments are the following:*

- $\Omega = \{$*Defective, Non-defective*$\}$ *if our experiment is to inspect a light bulb.*

  There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ *where* $\omega_1 =$ *Defective and* $\omega_2 =$ *Non-defective.*

- $\Omega = \{$*Heads, Tails*$\}$ *if our experiment is to note the outcome of a coin toss.*

  This time, $\Omega = \{\omega_1, \omega_2\}$ *where* $\omega_1 =$ *Heads and* $\omega_2 =$ *Tails.*

- *If our experiment is to roll a die then there are six outcomes corresponding to the number that shows on the top. For this experiment,* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

  Some examples of events are the set of odd numbered outcomes $A = \{1, 3, 5\}$*, and the set of even numbered outcomes* $B = \{2, 4, 6\}$.

  The simple events of Ω *are* $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$*, and* $\{6\}$.

The outcome of a random experiment is uncertain until it is performed and observed. Note that sample spaces need to reflect the problem in hand.

Definition 1.3. *A* **trial** *is a single performance of an experiment and it results in an outcome.*

Example 1.4. *Some standard examples of a trial are:*

- *A roll of a die.*
- *A toss of a coin.*
- *A release of a chaotic double pendulum.*

An experimenter often performs more than one trial. Repeated trials of an experiment forms the basis of science and engineering as the experimenter learns about the phenomenon by repeatedly performing the same mother experiment with possibly different outcomes. This repetition of trials in fact provides the very motivation for the definition of probability.

Definition 1.5. *An* **n-product experiment** *is obtained by repeatedly performing n trials of some experiment. The experiment that is repeated is called the "mother" experiment.*

Example 1.6 (Toss a coin n times)**.** *Suppose our experiment entails tossing a coin n times and recording* `H` *for Heads and* `T` *for Tails. When* $n = 3$*, one possible outcome of this experiment is* `HHT`*, ie. a Head followed by another Head and then a Tail. Seven other outcomes are possible.*

The sample space for "toss a coin three times" experiment is:

$$\Omega = \{\mathtt{H}, \mathtt{T}\}^3 = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{HTT}, \mathtt{THH}, \mathtt{THT}, \mathtt{TTH}, \mathtt{TTT}\} \ ,$$

with a particular sample point or outcome $\omega = \mathtt{HTH}$, and another distinct outcome $\omega' = \mathtt{HHH}$. An event, say A, that 'at least two Heads occur' is the following subset of Ω:

$$A = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{THH}\} \ .$$

Another event, say B, that 'no Heads occur' is:

$$B = \{\texttt{TTT}\}$$

Note that the event B is also an outcome or sample point. Another interesting event is the empty set $\emptyset \subset \Omega$. The event that 'nothing in the sample space occurs' is $\emptyset$.

EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
Ω	–	set of all outcomes of the experiment.
ω	–	an individual outcome in Ω, called a simple event.
$A \subseteq \Omega$	–	a subset A of Ω is an event.
Trial	–	one performance of an experiment resulting in 1 outcome.

1.2 Probability

The mathematical model for probability or the probability model is an axiomatic system that may be motivated by the intuitive idea of 'long-term relative frequency'. If the axioms and definitions are intuitively motivated, the probability model simply follows from the application of logic to these axioms and definitions. No attempt to define probability in the real world is made. However, the application of probability models to real-world problems through statistical experiments has a fruitful track record. In fact, you are here for exactly this reason.

Idea 1.7 (The long-term relative frequency (LTRF) idea)**.** *Suppose we are interested in the fairness of a coin, i.e. if landing Heads has the same "probability" as landing Tails. We can toss it n times and call $N(\texttt{H}, n)$ the fraction of times we observed Heads out of n tosses. Suppose that after conducting the tossing experiment* 1000 *times, we rarely observed Heads, e.g.* 9 *out of the* 1000 *tosses, then $N(\texttt{H}, 1000) = 9/1000 = 0.009$. Suppose we continued the number of tosses to a million and found that this number approached closer to* 0.1*, or, more generally, $N(\texttt{H}, n) \to 0.1$ as $n \to \infty$. We might, at least intuitively, think that the coin is unfair and has a lower "probability" of* 0.1 *of landing Heads. We might think that it is fair had we observed $N(\texttt{H}, n) \to 0.5$ as $n \to \infty$. Other crucial assumptions that we have made here are:*

1. **Something Happens***: Each time we toss a coin, we are certain to observe Heads* **or** *Tails, denoted by $\texttt{H} \cup \texttt{T}$. The probability that*

"something happens" is 1*. More formally:*

$$N(\mathtt{H} \cup \mathtt{T}, n) = \frac{n}{n} = 1.$$

This is an intuitively reasonable assumption that simply says that one of the possible outcomes is certain to occur, provided the coin is not so thick that it can land on or even roll along its circumference.

2. **Addition Rule***: Heads and Tails are mutually exclusive events in any given toss of a coin, i.e. they cannot occur simultaneously. The intersection of mutually exclusive events is the empty set and is denoted by* $\mathtt{H} \cap \mathtt{T} = \emptyset$*. The event* $\mathtt{H} \cup \mathtt{T}$*, namely that the event that "coin lands Heads* **or** *coin lands Tails" satisfies:*

$$N(\mathtt{H} \cup \mathtt{T}, n) = N(\mathtt{H}, n) + N(\mathtt{T}, n).$$

3. *The coin-tossing experiment is repeatedly performed in an* **independent** *manner, i.e. the outcome of any individual coin-toss does not affect that of another. This is an intuitively reasonable assumption since the coin has no memory and the coin is tossed identically each time.*

We will use the LTRF idea more generally to motivate a mathematical model of probability called probability model. Suppose A is an event associated with some experiment $\mathcal{E}$, so that A either does or does not occur when the experiment is performed. We want the probability that event A occurs in a specific performance of $\mathcal{E}$, denoted by $\mathbb{P}(A)$, to intuitively mean the following: if one were to perform a super-experiment $\mathcal{E}^\infty$ by independently repeating the experiment $\mathcal{E}$ and recording $N(A, n)$, the fraction of times A occurs in the first n performances of $\mathcal{E}$ within the super-experiment $\mathcal{E}^\infty$. Then the LTRF idea suggests:

$$N(A, n) := \frac{\text{Number of times } A \text{ occurs}}{n = \text{Number of performances of } \mathcal{E}} \to \mathbb{P}(A), \ as \quad n \to \infty \quad (1.1)$$

We first begin by defining certain collections of sets that will be the prototype for collections of events:

Definition 1.8 (Sigma algebras)**.** *Let* Ω *be a set: We say that a collection of subsets of* Ω*,* $\mathcal{F}$ *is a* **sigma-algebra/ sigma-field/** σ**-algebra** *if it satisfies the following properties:*

1. $\mathcal{F}$ *contains* Ω*, i.e.* $\Omega \in \mathcal{F}$*.*

2. *The collection* $\mathcal{F}$ *is closed under complementation*

$$A \in \mathcal{F} \implies A^C \in \mathcal{F}.$$

3. *The collection $\mathcal{F}$ is closed under countable unions*

$$A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_i A_i \in \mathcal{F}.$$

Remark 1.9. *For those not familiar with unions of a countable collection of sets, we define*

$$\bigcup_i A_i := \{\omega : \text{ there exists an } i \text{ such that } \omega \in A_i\}.$$

That is, $\omega \in \bigcup_i A_i$ if there is a set in the sequence $A_1, A_2, \ldots$ that contain ω.

Similarly we can define the countable intersection as

$$\bigcap_i A_i := \{\omega : \text{ for all } i \ \omega \in A_i\}.$$

That is, $\omega \in \bigcap_i A_i$ if it is in all $A_1, A_2, \ldots$.

Now, we are finally ready to define probability and events.

Definition 1.10 (Probability)**.** *Let $\mathcal{E}$ be an experiment with sample space Ω. Let $\mathcal{F}$ denote σ-algebra as in Definition 1.8. A* **probability measure** *is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying the following conditions:*

1. *The 'Something Happens' axiom holds, i.e. $\mathbb{P}(\Omega) = 1$.*
2. *The 'Addition Rule' axiom holds, i.e. for $A, B \in \mathcal{F}$:*

$$A \cap B = \emptyset \quad \implies \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \ .$$

We call elements of $\mathcal{F}$, events and we will call $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability triple**.

1.2.1 Consequences of our Definition of Probability

It is important to realize that we accept the 'addition rule' as an axiom in our mathematical definition of probability (or our probability model) and we do **not** prove this rule. However, the facts which are stated (with proofs) below, are logical consequences of our definition of probability:

Lemma 1.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, then*

1. *For any event $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*

2. *For any two events $A, B \in \mathcal{F}$, we have the* **inclusion-exclusion principle***:*

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

3. *From inclusion-exclusion principle we get* **Boole's inequality***: for any two events $A, B \in \mathcal{F}$*

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

An immediate consequence of 1 is: If $A = \Omega$ then $A^c = \Omega^c = \emptyset$ and $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$.

Proof. To prove 1 we proceed as follows

$$\overbrace{\mathbb{P}(A) + \mathbb{P}(A^c)}^{LHS} \underbrace{=}_{+ \text{ rule } \because A \cap A^c = \emptyset} \mathbb{P}(A \cup A^c) \underbrace{=}_{A \cup A^c = \Omega} \mathbb{P}(\Omega)$$

$$\underbrace{=}_{\because \ \mathbb{P}(\Omega)=1} \overbrace{1}^{RHS} \qquad \underbrace{\Longrightarrow}_{LHS-\mathbb{P}(A) \ \& \ RHS-\mathbb{P}(A)} \mathbb{P}(A^c)$$

$$= 1 - \mathbb{P}(A)$$

To prove 2 we note that since:

$$\begin{aligned} A = (A \setminus B) \cup (A \cap B) \quad &\text{and} \quad (A \setminus B) \cap (A \cap B) = \emptyset, \\ A \cup B = (A \setminus B) \cup B \quad &\text{and} \quad (A \setminus B) \cap B = \emptyset \end{aligned}$$

the addition rule implies that:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) \\ \mathbb{P}(A \cup B) &= \mathbb{P}(A \setminus B) + \mathbb{P}(B) \end{aligned}$$

Substituting the first equality above into the second, we get:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B)$$

Finally we note that Booles inequality, 3, follows immediately from 2 since $-\mathbb{P}(A \cap B) \leq 0$. □

These basic properties can then be iterated to obtain similar statements when there is more than 2 events.

Lemma 1.12. *(Extended properties) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, then*

1. *The inclusion-exclusion principle extends similarly to any n events $A_1, A_2, \ldots, A_n \in \mathcal{F}$ as follows:*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} \sum_{i<\cdots<n} \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right)$$

2. *Once again by the inclusion-exclusion principle, the Boole's inequality* **(Union bound)** *generalises to any n events $A_1, A_2, \ldots, A_n \in \mathcal{F}$ as follows:*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

3. *For a sequence of mutually disjoint events $A_1, A_2, A_3, \ldots, A_n \in \mathcal{F}$:*

$$A_i \cap A_j = \emptyset \quad \text{for any } i \neq j \quad \Longrightarrow \quad \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n).$$

Proof. For the proof of 1 and 2, see the counting argument in https://en.wikipedia.org/wiki/Inclusion%E2%80%93exclusion_principle if you are curious.

3 follows from 1 since all intersections are empty. □

We have formally defined the **probability model** specified by the **probability triple** $(\Omega, \mathcal{F}, \mathbb{P})$ that can be used to model an **experiment** $\mathcal{E}$.

Next, let us take a detour into how one might interpret it in the real world. The following is an adaptation from Williams D, *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press, 2001, which henceforth is abbreviated as WD2001.

Probability Model	**Real-world Interpretation**
Sample space Ω	Set of all outcomes of an experiment
Sample point ω	Possible outcome of an experiment
(No counterpart)	Actual outcome $\omega^\star$ of an experiment
Event A, a (suitable) subset of Ω	The real-world event corresponding to A occurs if and only if $\omega^\star \in A$
$\mathbb{P}(A)$, a number between 0 and 1	Probability that A will occur for an experiment yet to be performed

Events in Probability Model	**Real-world Interpretation**
Sample space Ω	The certain even 'something happens'
The $\emptyset$ of Ω	The impossible event 'nothing happens'
The intersection $A \cap B$	'Both A and B occur'
$A_1 \cap A_2 \cap \cdots \cap A_n$	'All of the events $A_1, A_2, \ldots, A_n$ occur simultaneously'
The union $A \cup B$	'At least one of A and B occurs'
$A_1 \cup A_2 \cup \cdots \cup A_n$	'At least one of the events $A_1, A_2, \ldots, A_n$ occurs'
A^c, the complement of A	'A does not occur'
$A \setminus B$	'A occurs, but B does not occur'
$A \subset B$	'If A occurs, then B must occur'

1.2.2 More on Sigma Algebras

Generally one encounters four types of sigma algebras (you will understand the last two types after taking more advanced courses in mathematics, so it is fine to understand the ideas intuitively for now!) and they are:

- When the sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is a finite set with k outcomes and $\mathbb{P}(\omega_i)$, the probability for each outcome $\omega_i \in \Omega$ is known, then one typically takes the sigma-algebra $\mathcal{F}$ to be the set of all subsets of Ω called the **power set** and denoted by 2^Ω. The probability of each event $A \in 2^\Omega$ can be obtained by adding the probabilities of the outcomes in A, i.e., $\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\omega_i)$. Clearly, 2^Ω is indeed a sigma-algebra and it contains $2^{\#\Omega}$ events in it.

- When the sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ is a countable set then one typically takes the sigma-algebra $\mathcal{F}$ to be the set of all subsets of Ω. Note that this is very similar to the case with finite Ω except now $\mathcal{F} = 2^\Omega$ could have uncountably many events in it.

- If $\Omega = \mathbb{R}^d$ for finite $d \in \{1, 2, 3, \ldots\}$ then the **Borel sigma-algebra** is the smallest sigma-algebra containing all **half-spaces**, i.e., sets of the form

$$\{x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d : x_1 \leq c_1, x_2 \leq c_2, \ldots, x_d \leq c_d\},$$

for any $c = (c_1, c_2, \ldots, c_d) \in \mathbb{R}^d$. When $d = 1$ the half-spaces are the half-lines $\{(-\infty, c] : c \in \mathbb{R}\}$ and when $d = 2$ the half-spaces are the south-west quadrants $\{(-\infty, c_1] \times (-\infty, c_2] : (c_1, c_2) \in \mathbb{R}^2\}$, etc. (Equivalently, the Borel sigma-algebra is the smallest sigma-algebra containing all open sets in $\mathbb{R}^d$).

- Given a finite set $\mathbb{S} = \{s_1, s_2, \ldots, s_k\}$, let Ω be the sequence space $\mathbb{S}^\infty := \mathbb{S} \times \mathbb{S} \times \mathbb{S} \times \cdots$, i.e., the set of sequences of infinite length that are made up of elements from $\mathbb{S}$. A set of the form

  $$A_1 \times A_2 \times \cdots \times A_n \times \mathbb{S} \times \mathbb{S} \times \cdots , \quad A_k \subset \mathbb{S} \text{ for all } k \in \{1, 2, \ldots, n\} ,$$

  is called a **cylinder set**. The set of events in $\mathbb{S}^\infty$ is the smallest sigma-algebra containing the cylinder sets.

1.3 Conditional Probability

Conditional probabilities arise when we have partial information about the result of an experiment which restricts the sample space to a range of outcomes. For example, if there has been a lot of recent seismic activity in Christchurch, then the probability that an already damaged building will collapse tomorrow is clearly higher than if there had been no recent seismic activity.

Conditional probabilities are often expressed in English by phrases such as:

"If A happens, what is the probability that B happens?"

or

"What is the probability that A happens if B happens?"

or

" What is the probability that A occurs given that B occurs?"

Definition 1.13 (Conditional Probability)**.** *Suppose we are given an experiment $\mathcal{E}$ with a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A, B \in \mathcal{F}$ (events), such that $\mathbb{P}(A) \neq 0$. Then, we define the* **conditional probability** *of B given A by,*

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} . \tag{1.2}$$

It turns out that the conditional probability is just a restriction of the events to A and it is as such, a probability measure.

Lemma 1.14. *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *then for* $A \in \mathcal{F}$ *with* $\mathbb{P}(A) \neq 0$,

$$\mathbb{P}(\cdot|A) : \mathcal{F} \to [0,1]$$

is a probability measure as in Definition 1.10 over $(\Omega, \mathcal{F})$.

Proof. Exercise! □

It is now clear from Lemma 1.14 that $(\Omega, \mathcal{F}, \mathbb{P}(\cdot|A))$ is a probability triple and as such Lemmas 1.11 and 1.12 holds. Hence, there is no distinction in how we can work with conditional probabilities versus regular probabilities.

1.3.1 Bayes' Theorem

Next we look at one of the most elegant applications of the definition of conditional probability along with the addition rule for a partition of Ω called *Bayes' Theorem*. We will present a two event case first called *Bayes' Rule* and then present the more general case of the Theorem.

This is useful because many problems involve reversing the order of conditional probabilities. Suppose we want to investigate some phenomenon A and have an observation B that is evidence about A: for example, A may be breast cancer and B may be a positive mammogram. Then Bayes' Theorem tells us how we should update our probability of A, given the new evidence B.

Or, put more simply, Bayes' Rule is useful when you know $P(B|A)$ but want $P(A|B)$!

Proposition 1.15 (Bayes' Rule)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple, let* $A, B \in \mathcal{F}$ *with* $\mathbb{P}(A), \mathbb{P}(B) > 0$, *then*

$$\mathbb{P}(A|B) \;=\; \frac{\mathbb{P}(A)\,\mathbb{P}(B|A)}{\mathbb{P}(B)} \;. \tag{1.3}$$

Proof. From the definition of conditional probability we have

$$\begin{aligned} \mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ \mathbb{P}(B \mid A) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \end{aligned} \tag{1.4}$$

From this we can see that

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

the first and last equality used (1.4) and in the second equality we just multiplied and divided by $\mathbb{P}(A)$. □

Before we see the more general form of Bayes' Rule, let us make a simple observation called the *total probaility theorem.*

Theorem 1.16 (Total probability)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple, suppose* $A_1 \cup A_2 \ldots \cup A_k \in \mathcal{F}$ *is a sequence of events with positive probability that partition the sample space, that is,* $A_1 \cup A_2 \cdots \cup A_k = \Omega$ *and* $A_i \cap A_j = \emptyset$ *for any* $i \neq j$*, then for some arbitrary event* $B \in \mathcal{F}$*.*

$$\mathbb{P}(B) \;=\; \sum_{h=1}^{k} \mathbb{P}(B \cap A_h) \;=\; \sum_{h=1}^{k} \mathbb{P}(B|A_h)\,\mathbb{P}(A_h) \tag{1.5}$$

Proof. Since $A_1, \ldots, A_k$ is a partition of Ω they are mutually exclusive (disjoint), hence

$$B \cap A_1, B \cap A_2, \ldots, B \cap A_k$$

are mutually exclusive. Thus the first equality follows from Lemma 1.12:3. The last equality follows from the definition of conditional probability. □

Reference to the Venn digram (you should draw below as done in lectures) will help you understand this idea for the four event case.

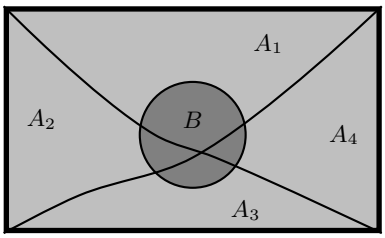

Figure 1.1: Reference to the Venn digram will help you understand this idea behind the proof of the total probability theorem in Theorem 1.16 for the four event case.

Theorem 1.17 (Bayes', 1763)**.** *Let everything be as in Theorem 1.16, and in addition assume that* $\mathbb{P}(B) > 0$*, then for any* $i = 1, \ldots, k$ *we have*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\,\mathbb{P}(A_j)} \tag{1.6}$$

Proof. From Definition 1.13, Lemmas 1.12 and 1.14, and Theorem 1.16 we

have

$$\begin{aligned}\mathbb{P}(A_h|B) &= \frac{\mathbb{P}(A_h \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A_h)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\mathbb{P}\left(\bigcup_{h=1}^{k}(B \cap A_h)\right)} = \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\sum_{h=1}^{k}\mathbb{P}\left(B \cap A_h\right)} \\ &= \frac{\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}{\sum_{h=1}^{k}\mathbb{P}(B|A_h)\,\mathbb{P}(A_h)}.\end{aligned}$$

□

It is customary to call $\mathbb{P}(A_h)$ the **prior probability of** A_h, i.e., before observing B or *a priori*, and $\mathbb{P}(A_h|B)$ the **posterior probability of** A_h, i.e., after observing B or *a posteriori*. Note that these names only make sense in the context of how you are modeling, in essence the above theorem does not differentiate between what is observed and not.

1.3.2 Independence and Dependence

In general, $P(A|B)$ and $P(A)$ are different, but sometimes the occurrence of B makes no difference, and gives no new information about the chances of A occurring. This is the idea behind independence. Events like "having blue eyes" and "having blond hair" are associated due to common genetic ancestry, but events like "my neighbour wins Lotto" and "I win Lotto" are not due to the Lotto machine being chaotically whirled around before ejection (as modelled by a well-stirred urn).

Definition 1.18 (Independence of two events)**.** *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, any two events* $A, B \in \mathcal{F}$ *are said to be* **independent** *if and only if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)\ . \tag{1.7}$$

Another way of making sense of the above definition is through the following lemma.

Lemma 1.19. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple, let* $A, B \in \mathcal{F}$ *be independent, then if* $P(B) > 0$ *we have*

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

Proof. From Definition 1.13 we have

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

□

The above lemma says that information about the occurrence of B does not affect the occurrence of A. If $\mathbb{P}(A) > 0$ we can use Lemma 1.19 with A, B reversed to get

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

which says that information about the occurrence of A does not affect the occurrence of B. So in a way, independence means that they have no effect on eachother.

If we have more than two events we can extend the notion of pairwise independence to an independent sequence.

Definition 1.20 (Independence of a sequence of events)**.** *Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, we say that a finite or infinite sequence of events $A_1, A_2, \ldots \in \mathcal{F}$ are independent if whenever $i_1, i_2, \ldots, i_k$ are distinct elements from the set of indices $\mathbb{N}$, then*

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k})$$

It should be noted that for a sequence of more than two events $A_1, A_2, \ldots \in \mathcal{F}$, pairwise independence (see Definition 1.18) is a weaker requirement than sequentially independent (see Definition 1.20). To make this clear, see the next example:

Example 1.21 (Pairwise independent events that are not jointly independent)**.** *Let a ball be drawn from an well-stirred urn containing four balls labelled 1,2,3,4. Consider the events $A = \{1, 2\}$, $B = \{1, 3\}$ and $C = \{1, 4\}$. Then,*

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(A)\,\mathbb{P}(B) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\
\mathbb{P}(A \cap C) &= \mathbb{P}(A)\,\mathbb{P}(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\
\mathbb{P}(B \cap C) &= \mathbb{P}(B)\,\mathbb{P}(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4},
\end{aligned}$$

but,

$$\frac{1}{4} = \mathbb{P}(\{1\}) = \mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) = \frac{2}{4} \times \frac{2}{4} \times \frac{2}{4} = \frac{1}{8} \ .$$

Therefore, inspite of being pairwise independent, the events A, B and C are not jointly independent.

1.4 Extension of probability*

Definition 1.10 is limited to only finite collections of sets, i.e. the additivity works for finitely many sets, which is certainly enough to obtain an understanding of probability. However, in the later stages we actually need the following extension of the definition of probability.

Definition 1.22 (Probability (Full))**.** *Let $\mathcal{E}$ be an experiment with sample space Ω. Let $\mathcal{F}$ denote σ-algebra as in Definition 1.8. A* **probability measure** *is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying the following conditions:*

1. *The 'Something Happens' axiom holds, i.e. $\mathbb{P}(\Omega) = 1$.*
2. *The 'Countably additive' axiom holds, i.e. let $\{E_i\}$ be a countable collection of events in $\mathcal{F}$ that are pairwise disjoint then:*

$$\mathbb{P}(\cup_i E_i) = \sum_i \mathbb{P}(E_i) \ .$$