INTRODUCTION

This course concerns two fields: probability and statistics.

Probability theory is a branch of mathematics which deals with random events. It originated as a mathematical theory of gambling and games of chance, and became a prominent part of mathematics in the 17th and 18th centuries. Scientists soon realised that the same concept of probability can be applied to many real-life phenomena, such as genetics, quantum physics, engineering reliability, and insurance, where there is an element of chance in the outcome. Probability was also applied in legal arguments, and is now an important part of the argument supporting DNA evidence in court. Some typical questions which can be answered using probability theory are:

- what are the chances that parents who both have Rh-positive blood type will have an Rh-positive child?
- what is the expected service lifetime of a power supply built with three redundant circuit breakers, if each circuit breaker has an expected life of 2 years?
- how fast would an epidemic of SARS spread in Australia, under reasonable assumptions?

Statistics is the science of drawing conclusions from data. The ‘data’ may be the results of a scientific experiment, some scientific observations, the results of an opinion poll, the medical records of a hospital, etc. Questions that can be answered using statistics include:

- what is the fair premium to charge for life insurance?
- do the results from our clinical trial show that the new drug is effective against AIDS?
- could the samples of bone from two archaeological sites have been deposited at the same time in prehistory?
- if three of a family’s children have died of sudden infant death syndrome, is this evidence that the parents are harming their children?

I view Statistics as a separate science, rather than a subset of mathematics, because the aim of statistics is to draw conclusions about the real world. To achieve this aim, statisticians use the mathematical theory of probability, as well as information technology (simulation, visualisation, statistical computing) and the scientific method.

The first $\frac{1}{3}$ of this unit is devoted to Probability Theory, and the remaining $\frac{2}{3}$ to Statistics.

1 PROBABILITY THEORY

1.1 Introduction

Originally, probability theory dealt with simple games of chance where all the possible outcomes are considered to be equally likely. For example if we roll an ordinary die, the numbers 1 to 6 are equally likely. There were several popular ways of reckoning the chances in a game:
the probability of winning

\[
\text{probability} = \frac{\text{number of favorable outcomes}}{\text{number of possible outcomes}}
\]

the odds of winning

\[
\text{odds} = \frac{\text{number of favorable outcomes}}{\text{number of unfavorable outcomes}}
\]

the expected value, or fair price to enter the game.

For example, suppose we roll a die, and win $3 if we roll a 4 or 6, otherwise we win nothing. There are two favorable outcomes (4 and 6). The probability of winning is

\[
\text{probability} = \frac{2}{6} = \frac{1}{3}
\]

while the odds of winning are

\[
\text{odds} = \frac{2}{4} = \frac{1}{2}
\]

or “1 to 2 against”. The fair price to enter the game is $1 (as we discuss below).

Initially it was not clear whether we should use probability, odds, or expected value to analyse games of chance. Also, once we go beyond the simple examples, it is not obvious how to calculate these quantities, and in some cases there was even confusion or disagreement about the right answer.

Example 1 (De Méré’s paradox). This problem confused some of the best mathematicians of the 17th century.

The Chevalier de Méré, a French nobleman, asked whether it is more likely to roll a six in four throws of a die, or to roll a double-six in 24 throws of two dice.

The chance of rolling a six in one throw of a single die is \(\frac{1}{6}\). So, de Méré argued, the chance of rolling a six in four throws should be \(4 \times \frac{1}{6} = \frac{2}{3}\). Similarly the chance of rolling a double-six in one throw of two dice is \(\frac{1}{36}\), and the chance of rolling a double-six in 24 throws should be \(24 \times \frac{1}{36} = \frac{2}{3}\). So these two events should be equally likely.

Other mathematicians disagreed. They argued that the chance of not getting a six in one throw of a die is \(\frac{5}{6}\), so the chance of getting no sixes in four throws is \(\frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = \left(\frac{5}{6}\right)^4\), so the chance of rolling at least one six in four throws is the complement of this, \(1 - \left(\frac{5}{6}\right)^4 = 0.52\). Similarly the chance of rolling at least one double-six in 24 throws of two dice is \(1 - \left(\frac{35}{36}\right)^{24} = 0.49\).

Who is right?

The flaw in De Méré’s reasoning is that the probability of at least one success in a series of trials cannot be obtained by adding up the chances of success at each trial. If this reasoning were correct, then in 6 throws of a die, the chance of getting a six would be \(6 \times \frac{1}{6} = 1\), that is, we would be certain to throw a six, while in 12 throws of a die, the chance of getting a six would be \(12 \times \frac{1}{6} = 2\) ?!?!?

De Méré’s paradox also shows the importance of stating a problem precisely. “Getting a six” in 4 throws is a vague description. Do we win a prize if we get at least one six in 4 throws, or do we win a prize each time we get a six?

Mathematicians quickly realised that, in order to avoid confusion in a probability problem, we need to answer two questions: (1) what are the possible outcomes of the game?, and (2) which of these outcomes are ‘favorable’ to the gambler?

This approach resolves De Méré’s paradox, for example. When we roll a die 4 times, the outcome is a sequence of four integers in the range 1 to 6. Any such sequence is possible: there are \(6^4 = 1296\) possible outcomes. An outcome is favorable to the gambler if there is at least one 6 in the sequence. Consider the unfavorable outcomes: these are the sequences which do not contain 6, in other words, sequences of four integers in the range 1 to 5. There are \(5^4\) unfavorable outcomes. Hence, there are
\[6^4 - 5^4 = 671\] favorable outcomes. Thus the probability of rolling at least one six in 4 throws of a die is \[671/1296 = 0.518\]. Similarly if we roll a pair of dice 24 times, there are \[36^{24} = 2.25 \times 10^{37}\] possible outcomes, and \[36^{24} - 35^{24} = 1.10 \times 10^{37}\] favorable outcomes, so the probability of rolling at least one double-six in 24 throws of a pair of dice is \[1.10/2.25 = 0.491\]. So De Méré was wrong.\(^1\)

**Example 2 (Birthday problem).** "What an incredible coincidence! Two of the students in my class of 30 share the same birthday. That is very unlikely, considering the chance that two people have the same birthday is \[1/365\].”

The flaw in this reasoning is that we have not carefully defined the problem. The real question is to find the probability that, in a class of 30 people, there are (at least) two people in the class with the same birthday. While it is reasonable to calculate that there is a chance of \[1/365\] that a given pair of students share the same birthday, the real question is to find the probability that any pair of students in the class of 30 share a birthday.

Each student’s birthday could be any one of the 365 days of the year, so there are \[365^{30} = 7.4 \times 10^{76}\] possible outcomes. The easiest way to calculate the probability is to consider the complementary event, that no-one in the class shares the same birthday, that is, everyone has a different birthday. The number of ways of allocating different birthdays to these 30 students is \[365 \cdot 364 \cdot \ldots \cdot 336 = 2.1 \times 10^{76}\]. So the probability that everyone has a different birthday is

\[
\frac{365 \cdot 364 \cdot \ldots \cdot 336}{365 \cdot 365 \cdot \ldots \cdot 365} = 0.294.
\]

So the probability that at least two students will share a birthday, is \[1 - 0.294 = 0.706\]. That is, in a class of 30 people, there is a 71% chance that two students will share a birthday. Similarly in a class of 40 people, the probability is over 90%.

This example again shows how important it is to clarify what are the possible outcomes of the experiment, and which of the outcomes are favorable.

The example also shows that probability sometimes runs counter to our intuition. Why is the probability of a ‘coincidence’ so high? A short answer is that there are many pairs of students in the class (435 pairs in a class of 30) so there are many more opportunities for a ‘coincidence’ than we imagine.

Examples like these showed the need for a **set of agreed rules** or axioms, for calculating the probabilities of combinations of events. Eventually these rules developed into a rigorous mathematical theory of probability.

Why do we need a mathematical theory of probability? At school we all learned to do elementary probability problems using intuition and pictures (Venn diagrams, tree diagrams etc). I want to convince you that these intuitive methods are not adequate in the grown-up world. They often do not work on complex problems, and they can produce wrong answers if we apply them inappropriately. To be sure of our answers, we need to formulate probability problems in a rigorous way.

Here is another example which confuses many professional mathematicians in the 21st century.

**Example 3 (Monty Hall problem).** In a game show, there are three doors. The host of the show, Monty Hall, chooses one of the doors at random, and hides the prize behind that door. Only the host knows where the prize is.

Then the contestant is asked to choose one of the three doors. Instead of opening this door, the host opens another door that he knows does not conceal the prize. (That is, the host opens a door that does not conceal the prize and was not chosen by the contestant).

One of the two unopened doors conceals the prize. The host offers the contestant the choice of opening either door: the door originally chosen, or the other unopened door.

---

\(^1\)However, De Méré’s calculations do have a meaning. They give us the **expected number** of successes, rather than the probability of **at least one** success. The expected number of sixes obtained in four throws of a die is \[4 \times 1/6 = 2/3\] and the expected number of double-sixes in 24 throws of two dice is \[24 \times 1/36 = 2/3\]. So de Méré’s paradox can also be interpreted as arising from confusion between the **probability** of success and the **expected number** of successes.
Should the contestant stick with the door they chose first, or switch to the other door?

[Can you solve this with pictures and tree diagrams? If so, are you sure the answer is right?]

Apart from its tendency to be imprecise, there are several other weaknesses in the classical approach to probability:

- it relies on counting the outcomes, which could be troublesome in complicated problems.
- it assumes the outcomes are equally likely. So it can’t be used to study genetics or life insurance where the possible outcomes have unequal chances.
- it assumes there are finitely many possible outcomes. So it can’t be used to study phenomena that occur in continuous time (like the decay of radioactive elements).

In modern probability theory, we can deal with all these issues easily.

1.2 Axioms of probability

Reference: Rice (1995), Chapter 1, Section 1.3

The great Russian mathematician Andrei Nikolayevich Kolmogorov laid out in 1933 a set of rules for probability models which is now widely accepted.

1.2.1 Sample space

To describe a game of chance or any other random experiment, we first specify the set \( \Omega \) of all possible outcomes. This is sometimes called the sample space or probability space. When the game is played, or the experiment is performed, the actual outcome will be one element \( \omega \in \Omega \).

It is often vitally important to specify \( \Omega \) clearly (examples will be given later). Here are some examples of sample spaces.

- When we toss a coin, the sample space is \( \Omega = \{ \text{Heads}, \text{Tails} \} \).
- When we roll a six-sided die, the sample space is \( \Omega = \{1, 2, 3, 4, 5, 6\} \).
- In classical Mendelian genetics, a person inherits two copies of a gene, one from each parent. The gene has two variants (‘alleles’) which we may call \( A \) and \( a \). Thus a person’s genetic makeup may be \( AA, Aa, aA, \text{ or } aa \) where the first and second letters denote the genes inherited from the mother and father respectively. So the sample space for a person’s genotype (for this gene) is

\[
\Omega = \{ AA, Aa, aA, aa \}.
\]

Typically \( Aa \) and \( aA \) are indistinguishable so we might say

\[
\Omega = \{ AA, Aa, aa \}.
\]

- Suppose we run a machine until it fails, and record the elapsed time \( t \) after which failure occurred (the ‘failure time’). This could be any positive number \( t > 0 \), so the sample space is the set of all positive numbers,

\[
\Omega = \{ t : t > 0 \} = (0, \infty).
\]

Notice that \( \Omega \) is always a set. The elements of the set are the possible outcomes.

If an experiment or a game involves a sequence of several steps, then each possible outcome of the experiment is a sequence consisting of the outcomes of each step.
• When we toss a coin twice, the possible outcomes are \((H,H), (H,T), (T,H)\) and \((T,T)\) where for example \((H,T)\) indicates that the first toss was Heads and the second was Tails. The sample space
\[
\Omega = \{(H,H), (H,T), (T,H), (T,T)\}
\]
consists of all possible pairs of the letters H and T, so it can also be described as the Cartesian product
\[
\Omega = \{H,T\} \times \{H,T\}.
\]
• When we roll a die three times, the possible outcomes are all the possible sequences of three integers between 1 and 6, so the sample space is
\[
\Omega = \{(1,1,1), (1,1,2), \ldots, (6,6,6)\} = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} = \{1,2,3,4,5,6\}^3,
\]
the set of all triples of integers 1 to 6.

**Exercise 1.** In De Méré’s paradox, what is the sample space \(\Omega\) for the experiment in which we roll a die 4 times? What about rolling a pair of dice 24 times?

### 1.2.2 Events

A random *event* is something which may or may not occur, depending on the outcome of the experiment. It is mathematically represented as a subset \(A \subseteq \Omega\) (consisting of all the outcomes \(\omega\) which are favorable to this event). The event \(A\) is said to “occur” if the actual outcome \(\omega\) belongs to \(A\) \((\omega \in A)\) and “does not occur” if \(\omega \notin A\). We shall often write events using set braces, writing \(\{\ldots\}\) as an abbreviation for \(\{\omega \in \Omega : \ldots \text{ holds}\}\). Here is a “translation table” between the language of set theory and the language of events.

<table>
<thead>
<tr>
<th>Set notation</th>
<th>Events language</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A \cap B)</td>
<td>both (A) and (B) occur</td>
</tr>
<tr>
<td>(A \cup B)</td>
<td>either (A) or (B) (or both) occur</td>
</tr>
<tr>
<td>(A^c)</td>
<td>(A) does not occur</td>
</tr>
<tr>
<td>(A \cap B = \emptyset)</td>
<td>events (A) and (B) are disjoint (cannot occur together)</td>
</tr>
<tr>
<td>(A = \Omega)</td>
<td>event (A) always occurs</td>
</tr>
<tr>
<td>(A = \emptyset)</td>
<td>event (A) never occurs</td>
</tr>
</tbody>
</table>

Here \(A^c\) denotes the complement of \(A\), and \(\emptyset\) is the empty set.

**Example 4.** Suppose we toss a coin twice. The sample space is \(\Omega = \{HH, HT, TH, TT\}\). Here is a translation table of some common events.

<table>
<thead>
<tr>
<th>Events language</th>
<th>Set notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>{first toss is Heads}</td>
<td>(HH, HT)</td>
</tr>
<tr>
<td>{first toss is Tails}</td>
<td>(TH, TT)</td>
</tr>
<tr>
<td>{second toss is Tails}</td>
<td>(HT, TT)</td>
</tr>
<tr>
<td>{at least one toss is Heads}</td>
<td>(HH, HT, TH)</td>
</tr>
</tbody>
</table>

**Exercise 2.** In De Méré’s paradox, when we roll a die 4 times, what is the event that we get at least one six? (Identify this event as a subset of \(\Omega\)).
1.2.3 Probabilities

Finally we must assign a probability to each event. If $A$ is an event (a subset of $\Omega$), then $P(A)$ will denote the probability of $A$.

**Example 5 (Fair coin toss).** If we toss a coin, the sample space is $\Omega = \{H, T\}$. The usual assumptions are that Heads and Tails are equally likely. We must assign a probability $P(A)$ to each subset $A \subseteq \Omega$; so the appropriate values are

\[
\begin{align*}
P(\emptyset) &= 0 \\
P(\{H\}) &= \frac{1}{2} \\
P(\{T\}) &= \frac{1}{2} \\
P(\{H, T\}) &= 1.
\end{align*}
\]

In classical probability theory, it was typically assumed that each possible outcome is equally likely. But this is not always true in real life. We often need to study situations where the probabilities of different outcomes are unequal.

**Example 6 (Biased coin).** Here is a probability model for a biased coin, where the probability of tossing Heads is 0.6 and the probability of Tails is 0.4. The sample space is $\Omega = \{H, T\}$.

The probabilities of each event are:

\[
\begin{align*}
P(\emptyset) &= 0 \\
P(\{H\}) &= 0.6 \\
P(\{T\}) &= 0.4 \\
P(\{H, T\}) &= 1
\end{align*}
\]

Instead of prescribing the values of the probabilities $P(A)$, we shall just assume that the values $P(A)$ have been specified somehow (perhaps by performing experiments, or by making some assumptions, etc). Kolmogorov’s axioms stipulate a set of basic rules which these probabilities should obey.

**Definition 1 (Kolmogorov’s Axioms).** A probability measure $P$ is a function which specifies a probability $P(A)$ for each event $A \subseteq \Omega$ and which satisfies the following:

- (P1) $P(\Omega) = 1$;
- (P2) $P(A) \geq 0$ for all $A \subseteq \Omega$;
- (P3) if $A$ and $B$ are disjoint ($A \cap B = \emptyset$) then $P(A \cup B) = P(A) + P(B)$.

In the mathematical theory of probability we just assume that the probabilities $P(A)$ have been assigned so that these three axioms (P1)–(P3) hold. Then we derive logical consequences of these axioms. This gives a very powerful theory, because it applies to many different situations. We only have to check that the axioms (P1)–(P3) hold, and then the entire technology of probability theory is at our disposal.

In each application of this theory, we have to define the probabilities $P(A)$ ourselves, and then check that the axioms do hold.

**Example 7 (Fair coin toss).** Consider the probabilities $P(A)$ defined in Example 5. Clearly axioms (P1) and (P2) are true for this function $P$. It is not hard to check that axiom (P3) also holds by considering each pair of events $A, B$. So this is a probability measure.
Example 8 (Biased coin). It is just as easy to check that the model of a biased coin (Example 6) is also a probability measure.

Example 9 (Roll of a die). We can formalise the classical model of rolling a die as follows. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The usual assumptions are that each of the six outcomes is equally likely. So, for each event $A \subseteq \Omega$, define the probability

$$P(A) = \frac{\text{number of elements in } A}{6}.$$  

(Notice that we use an equation to define the probability of each event $A$, so that we avoid having to write down all the $2^6 = 64$ possible subsets.)

It is easy to check that axioms (P1)--(P3) hold for the particular function $P$ defined in Example 9. Axiom (P1) is trivial to verify, since there are 6 elements in $\Omega$. Axiom (P2) holds because the number of elements in any subset of $\Omega$ is a nonnegative number. To verify (P3) we just notice that if $A$ and $B$ are disjoint then

$$\text{(number of elements in } A \cup B) = (\text{number of elements in } A) + (\text{number of elements in } B)$$

so that $P(A \cup B) = P(A) + P(B)$. So axioms (P1)--(P3) hold for this function $P$.

In classical probability theory it was usually assumed that all possible outcomes are equally likely, as we just did in Example 9. In this case the axioms (P1)--(P3) are always satisfied, as we can see from the following theorem.

Theorem 1. Let $\Omega$ be a finite set. Suppose we define, for each $A \subseteq \Omega$,

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } \Omega}.$$  

Then $P$ is a probability measure.

Example 10. We toss a coin twice. The sample space $\Omega$ is $\Omega = \{\text{HH, HT, TH, TT}\}$. If we believe the coin is fair, then the appropriate probability measure $P$ is

$$P(A) = \frac{\text{number of elements in } A}{4}$$

and the Theorem above says that this is a probability measure.

In the examples above, we had to define the value of $P(A)$ for every subset $A$. That could be a lot of work in a complex problem. Fortunately, if the sample space $\Omega$ is finite, then it is enough just to specify the probability of each outcome, rather than the probability of every event. In the example of a coin toss, we only have to specify the probability of Heads and the probability of Tails.

Theorem 2. Suppose $\Omega$ is a finite set, say $\Omega = \{1, 2, \ldots, m\}$. Let $p_1, \ldots, p_m$ be any given nonnegative numbers ($p_i \geq 0$ for all $i$) which sum to 1 ($\sum_{i=1}^{m} p_i = 1$). Define

$$P(A) = \sum_{i \in A} p_i$$

for each $A \subseteq \Omega$. Then $P$ is a probability measure. It satisfies $P(\{i\}) = p_i$ for each possible outcome $i$.

Example 11. In classical Mendelian genetics, for a single gene with two alleles, the sample space is $\Omega = \{\text{AA, Aa, aa}\}$.
Mendel assigned probabilities

\[
\begin{align*}
P(\{AA\}) &= p^2 \\
P(\{Aa\}) &= 2pq \\
P(\{aa\}) &= q^2
\end{align*}
\]

where \( p \) is the population frequency of the allele \( A \), and \( q = 1 - p \) is the frequency of the allele \( a \). This is enough information to define a probability measure, using the theorem above. For example we get

\[
P(\{AA, Aa\}) = P(\{AA\}) + P(\{Aa\}) = p^2 + 2pq.
\]

Probability measures may also be defined when \( \Omega \) is infinite, but they require more work.

**Example 12 (Radioactive decay).** Each atom of a radioisotope eventually decays, emitting energy and breaking down into other atoms. The elapsed time \( t \) in seconds from the creation of the atom until decay occurs can be any positive real number \( t > 0 \). The set of possible outcomes is

\[
\Omega = \{ t : t > 0 \} = (0, \infty),
\]

the positive half line. The event that the particle decays within the first 10 seconds is \( A = (0, 10] \). The event that it survives longer than 60 seconds is \( A = (60, \infty) \).

Models from quantum physics tell us that the probability that the atom will **not** decay during the first \( s \) seconds is

\[
P(\{ t : t > s \}) = P((s, \infty)) = e^{-\lambda s}
\]

where \( \lambda \) is a known constant (the ‘rate constant’). Clearly this implies that the probability the atom will **decay** during the first \( s \) seconds is

\[
P(\{ t : t \leq s \}) = P([0, s]) = 1 - e^{-\lambda s}.
\]

It turns out that this is enough information to determine the probability \( P(A) \) of essentially any subset \( A \subseteq [0, \infty) \).

### 1.2.4 Improperly specified probabilities

To convince you that it is important to set up a probability measure carefully, here are some real-world examples.

**Example 13 (Envelope problem).** A game show host offers the contestant a choice of two envelopes. The contestant chooses one at random with equal probability, and opens it. It contains a cheque for \$1 million. The host then says that the other envelope contains either a cheque for twice this amount or half this amount, and offers the contestant the choice of either one of the two envelopes. What should the contestant choose?

The two envelopes contain either \$1 million and \$2 million, or \$1 million and \$0.5 million. No matter how we specify the sample space \( \Omega \), there is a problem in specifying the probability measure \( P \). Let \( A \) be the event

\[
A = \{ \text{the two envelopes contain \$1 million and \$2 million} \}.
\]

There is no information in the statement of the problem that allows us to determine \( P(A) \). The host puts the cheques in the envelopes, with an unspecified probability of inserting \$3 million as opposed to \$1.5 million. One of the envelopes is subsequently chosen at random with equal probability, but this is not relevant to the probability of \( A \). So this problem is not a well-defined probability model; \( P(A) \) is undefined; and there is no solution.
Example 14 (Prosecutor’s Fallacy). A person is arrested for murder. DNA from the crime scene matches the DNA of the accused. At the trial, a geneticist testifies that there is a chance of 1 in 80 million that a randomly-selected person would have a DNA profile that matches the profile from the crime scene. The prosecutor tells the jury that this means “the chance that the accused is innocent is 1 in 80 million”.

This conclusion is a fallacy. Indeed it arises so often in court cases that it is widely known as the Prosecutor’s Fallacy. Assuming the geneticist’s testimony is correct, what we know is that a randomly-chosen person would have a chance of $1/80,000,000$ of matching the DNA at the crime scene. The geneticist has a probability model in which the possible outcomes are the people in some population, and it is known that 1/80,000,000 of people in this population match the DNA at the crime scene.

This emphatically does not tell us about the probability that the accused is guilty. The accused is not a randomly-chosen person! For example, if the murderer is believed to be of Albanian descent (say) and the police arrest someone of Albanian descent, there is a much higher chance of a DNA match. The geneticist could be asked to provide the probability that a randomly-chosen person of Albanian descent would match the DNA at the crime scene. But even this is not the right answer.

To determine the chance that a person accused of murder would match the DNA at the crime scene, we need to know how the police selected the accused. The event that a person is accused of murder” is not part of the geneticist’s probability model. This kind of information is not even part of the science of genetics, and without further information or assumptions, genetic evidence on its own is insufficient to pronounce on the guilt of the accused.

This argument has been a feature of many recent court cases, including a case that was reported prominently in the West Australian newspaper in 2004.

1.3 Consequences of the axioms

The methods of modern probability theory do not assume that $P$ has a particular form; for example they do not assume that all outcomes are equally likely. They only assume that the axioms (P1)–(P3) hold.

This means that, whenever we can set up a probability model (by defining $\Omega$ and $P$) which satisfies the axioms (P1)–(P3), the model also satisfies all the statements of probability theory derived from these axioms.

Here are some consequences of the axioms (P1)–(P3).

Theorem 3. Suppose $P$ is a probability measure. Then

1. $P(\emptyset) = 0$;
2. $P(A) \leq 1$ for all $A \subseteq \Omega$;
3. $P(A^c) = 1 - P(A)$ for all $A \subseteq \Omega$;
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events $A, B \subseteq \Omega$;
5. $P(A \setminus B) = P(A) - P(A \cap B)$ for any events $A, B \subseteq \Omega$, where $A \setminus B = A \cap B^c$;
6. if $A \subseteq B$ then $P(A) \leq P(B)$.

For example, let us prove statement number 3, assuming only the axioms (P1)–(P3). Let $B = A^c$. Then $A \cap B = A \cap A^c = \emptyset$. Since $A$ and $B$ are disjoint, we may apply (P3) and conclude that $P(A \cup B) = P(A) + P(B)$. But $A \cup B = A \cup A^c = \Omega$, so (P1) implies that $P(A \cup B) = P(\Omega) = 1$. So we have $1 = P(A) + P(B) = P(A) + P(A^c)$, yielding $P(A^c) = 1 - P(A)$.

The complement rule $P(A^c) = 1 - P(A)$ is very useful in practice.
Example 15 (Monty Hall problem, revisited). Consider Example 3 again. There are several ways of specifying the sample space Ω for this problem, which we will not go into. Let A be the event that the first door chosen by the contestant contains the prize. Since the prize was placed at random with equal probability for each of the three doors, clearly \( P(A) = \frac{1}{3} \). Hence the probability of winning the prize if we stick with the original door is \( \frac{1}{3} \). Now the prize is either under the door we originally chose, or under the remaining unopened door. Hence the probability of winning the prize if we switch to the other door is \( P(A^c) = 1 - P(A) = 2/3 \). So it is better to switch.

Statement number 4 shows that De Méré’s reasoning (Example 1) was incorrect. Suppose \( A \) is the event that we roll a six on the first throw of a die, and \( B \) is the event that we roll a six on the second throw. Then the probability that we roll at least one six in the two throws is \( P(A \cup B) \), which is not equal to \( P(A) + P(B) \) but to \( P(A) + P(B) - P(A \cap B) \).

Exercise 3. Prove the other statements in Theorem 3.

1.4 Conditional probability and independence

Example 16. A standard mix of jellybeans contains raspberry (30%), strawberry (20%), blueberry (20%), coconut (10%) and licorice (10%). Suppose we pick a jellybean at random without looking, and find that it is red (i.e. either raspberry or strawberry). What is the probability that, when we eat it, we will find that it is raspberry?

Think of the red jellybeans as a subset or subpopulation of all jellybeans. Red jellybeans constitute 30% + 20% = 50% of the entire population of jellybeans. The probabilities of the different flavours in this subpopulation are different from the original probabilities in the entire population listed above. Raspberry jellybeans make up 30% of the entire population, but they make up

\[
\frac{30}{50} = \frac{3}{5}
\]

of the subpopulation of red jellybeans. So, given that a jellybean is red, the probability that it is raspberry is 3/5 or 60%. This is an example of a conditional probability.

Definition 2. Suppose \( A \) and \( B \) are two events in the sample space \( \Omega \) and that \( P(B) > 0 \). The conditional probability of \( A \) given \( B \) is

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}.\]

Often we have some partial information about the outcome of a game or experiment. In card games we often know the values of some of the cards, but not all. A doctor may know the results of some pathology tests on a patient, but does not know exactly what illness the patient has. Conditional probabilities are used in this situation.

Example 17. At the casino, the banker rolls a die (Example 9) without showing us the outcome. We win if the outcome is at least 4. The banker tells us that the outcome was an even number. What is the chance that we win, given this information?

Let \( A = \{ \text{win} \} = \{4, 5, 6\} \) and \( B = \{ \text{roll an even number} \} = \{2, 4, 6\} \). Then \( P(A \cap B) = P(\{4, 6\}) = 2/6 = 1/3 \) and \( P(B) = 3/6 = 1/2 \), so the conditional probability of \( A \) given \( B \) is

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}.
\]

Theorem 4. Suppose \( P \) is a probability measure. Let \( B \) be an event with \( P(B) > 0 \). Then the function

\[
Q(A) = P(A \mid B)
\]

is a probability measure on \( \Omega \).
You can check that this is true. For example, to prove that (P1) holds, just write

\[
Q(\Omega) = P(\Omega \mid B) \\
= \frac{P(\Omega \cap B)}{P(B)} \text{ by definition} \\
= \frac{P(B)}{P(B)} \\
= 1.
\]

This theorem is very useful because it means we do not need any extra effort to prove statements about conditional probability. For example we know immediately that

\[
P(A^c \mid B) = 1 - P(A \mid B)
\]

using Theorem 3, part 3.

**Exercise 4.** If \( A, B, C \) are events with \( P(C) > 0 \) and \( A \subseteq B \), prove that \( P(A \mid C) \leq P(B \mid C) \).

Here are some extra facts that relate the conditional and unconditional probabilities.

**Theorem 5 (Multiplication law).** Suppose \( A \) and \( B \) are events in the sample space \( \Omega \) and \( P(B) > 0 \). Then

\[
P(A \cap B) = P(A \mid B) P(B).
\]

The multiplication law is just a rearrangement of the definition of conditional probability but it is useful to remember.

**Theorem 6 (Law of total probability).** Suppose \( A \) and \( B \) are events in the sample space \( \Omega \) such that \( P(B) > 0 \) and \( P(B^c) > 0 \). Then

\[
P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).
\]

To prove this we note that

\[
A = (A \cap B) \cup (A \cap B^c)
\]

(it may help to draw a Venn diagram). Also the two events \( A \cap B \) and \( A \cap B^c \) are disjoint because

\[
(A \cap B) \cap (A \cap B^c) = A \cap B \cap B^c = \emptyset.
\]

So by axiom (P3),

\[
P(A) = P(A \cap B) + P(A \cap B^c).
\]

Then we use the multiplication law to get the result.

**Example 18.** To drive to the city of Perth from UWA campus, I can either go along Thomas Street or Mounts Bay Road. Half the time I choose Thomas St. On Thomas Street I have a 1% chance of being clocked by a police radar, while on Mounts Bay Rd the chance is 2%. On an average day what is the chance that I will be clocked by the radar?

Let \( A = \{ \text{clocked by radar} \} \) and \( B = \{ \text{drive along Thomas St} \} \). Then we have \( P(B) = 0.5 \) while \( P(A \mid B) = 0.01 \) and \( P(A \mid B^c) = 0.02 \). So

\[
P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c) \\
= 0.01 \times 0.5 + 0.02 \times 0.5 \\
= 0.015
\]

or 1.5%. 

11
Definition 3. Events $A$ and $B$ are independent if

$$P(A \cap B) = P(A) P(B).$$

Assuming $P(B) > 0$, then $A$ and $B$ are independent if and only if

$$P(A \mid B) = P(A)$$

i.e. the conditional probability that $A$ occurs, given $B$, is the same as the (unconditional) probability of $A$. This says that knowing whether $B$ occurred does not influence the chance that $A$ occurs.

Example 19. Suppose we roll a die and all six outcomes are equally likely (see Example 9). Consider the events $A = \{\text{even number}\} = \{2, 4, 6\}$ and $B = \{\text{number at least 3}\} = \{3, 4, 5, 6\}$. We have

$$P(A \mid B) = \frac{P\{4, 6\}}{P\{3, 4, 5, 6\}} = \frac{1}{2}$$

$$P(A) = P\{2, 4, 6\} = \frac{1}{2}$$

so $A$ and $B$ are independent.

Example 20 (Successive tosses of a coin). Suppose we toss a coin twice, and all possible outcomes are equally likely (see Example 10). Let $A$ be the event that the first toss is Heads, and $B$ the event that the second toss is Heads. We have

$$A = \{\text{first toss is H}\} = \{HH, HT\}$$

$$B = \{\text{second toss is H}\} = \{HH, TH\}$$

so

$$P(A) = P\{HH, HT\} = \frac{1}{2}$$

$$P(B) = P\{HH, TH\} = \frac{1}{2}$$

$$P(A \cap B) = P\{HH\} = \frac{1}{4}$$

Since

$$P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(A \cap B)$$

the events $A$ and $B$ are independent.

Don’t get confused between “independent” and “disjoint” events! Disjoint events exclude each other, and are usually not independent.

Independence is a property related to the probabilities of the events, so it depends on the probability measure $P$. It is unfortunately very common to make the mistake of assuming that two events are independent without justification.

Example 21 (Cot deaths). A senior paediatrician’s misleading testimony led to a mother being wrongfully jailed for the murder of her two baby sons. Professor Sir Roy Meadow stated at the trial of Mrs Sally Clark in 1999 that there was just a “one in 73 million” chance that two babies from an affluent family like hers could suffer cot death.

The basis of Professor Meadow’s reasoning was apparently that, since cot deaths occur at a rate of 1.17 deaths per 10,000 babies in this population, the chance of two children dying of cot death is $0.000117^2 = 0.0000000136$ or about 1 in 73 million. This calculation assumes that the cot deaths of different children are independent.
The Royal Statistical Society objected to this testimony. There is no basis for believing that two cot deaths in a family are independent; for example it might be that there are genetic or environmental causes of cot death, common to all babies in certain families. In fact the available statistics suggest that multiple cot deaths are not independent.

Mrs Clark’s conviction, and the convictions of three other women, were later overturned. The UK’s General Medical Council found Professor Meadow guilty of serious professional misconduct for giving evidence beyond his expertise at Mrs Clark’s trial. Prof Meadow is now barred from practising medicine in the UK.

**Example 22 (Challenger Accident).** In 1987 the space shuttle **Challenger** exploded after launch, with the loss of 7 lives and hundreds of millions of dollars. The famous physicist Richard Feynman was part of the team investigating the accident. His report begins:

> It appears that there are enormous differences of opinion as to the probability of a failure with loss of vehicle and of human life. The estimates range from roughly 1 in 100 to 1 in 100,000. The higher figures come from the working engineers, and the very low figures from management. What are the causes and consequences of this lack of agreement? Since 1 part in 100,000 would imply that one could put a Shuttle up each day for 300 years expecting to lose only one, we could properly ask “What is the cause of management’s fantastic faith in the machinery?”

The management’s justification was that, for a failure to occur, several things must go wrong at the same time. For example if there are three things that must fail before an accident occurs, each with a risk of 1/100, then the chance of them all failing together is (1/100)^3 = 1/1000000. But this assumes that the three things are independent; and there is no justification for this assumption.

There is a further complication about the independence of several events.

**Definition 4.** Three events \(A, B, C\) are called mutually independent if all of the following statements are true:

- \(P(A \cap B) = P(A)P(B)\);
- \(P(A \cap C) = P(A)P(C)\);
- \(P(B \cap C) = P(B)P(C)\); and
- \(P(A \cap B \cap C) = P(A)P(B)P(C)\).

The first line above states that \(A\) and \(B\) are independent in the sense of Definition 3. The second line says that \(A\) and \(C\) are independent, and the third line that \(B\) and \(C\) are independent. But the last line is something new, since it involves all three of the events.

**Example 23.** Suppose we toss a coin three times, and each of the 8 possible outcomes is equally likely. Let

\[
A = \{\text{first toss is Heads}\} \\
B = \{\text{second toss is Heads}\} \\
C = \{\text{third toss is Heads}\}.
\]

Then we have \(P(A) = P(B) = P(C) = 1/2\) and \(P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4\) and \(P(A \cap B \cap C) = 1/8\). It follows that \(A, B, C\) are mutually independent.
**Example 24.** Consider the model where we toss a fair coin twice (Example 10). Let

\[ A = \{ \text{first toss is Heads} \} = \{ HH, HT \} \]
\[ B = \{ \text{second toss is Heads} \} = \{ HH, TH \} \]
\[ C = \{ \text{first and second tosses are different} \} = \{ HT, TH \}. \]

Then we have \( P(A) = P(B) = P(C) = 1/2 \) and \( P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4 \) so that \( A \) and \( B \) are independent, \( B \) and \( C \) are independent, and \( A \) and \( C \) are independent. However \( A \cap B \cap C = \emptyset \) so that \( P(A \cap B \cap C) = 0 \), so that \( P(A \cap B \cap C) \neq P(A)P(B)P(C) \). The three events \( A, B, C \) are not mutually independent.

So we see that it is possible for three events \( A, B, C \) to be “pairwise independent” (in the sense of Definition 3) but not mutually independent.

**Exercise 5.** We toss a coin three times; all 8 possible outcomes are equally likely. Are the following three events mutually independent?

\[ A = \{ \text{first toss is Heads} \} \]
\[ B = \{ \text{second toss is same as first toss} \} \]
\[ C = \{ \text{third toss is same as first toss} \} \]

When there are more than 3 events in question, here is the definition of mutual independence.

**Definition 5.** Events \( A_1, A_2, \ldots, A_n \) are called mutually independent if

- \( P(A_i \cap A_j) = P(A_i)P(A_j) \) for all \( i, j \) with \( 1 \leq i < j \leq n \);
- \( P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k) \) for any indices \( i, j, k \) satisfying \( 1 \leq i < j < k \leq n \);
- in general,
  \[ P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_m}) \]
  for any indices \( i_1, \ldots, i_m \) satisfying \( 1 \leq i_1 < i_2 < \ldots < i_m \leq n \).

In particular, if \( A_1, \ldots, A_n \) are mutually independent, then

\[ P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n). \]