I. Introduction to Probability

Basic probability models

A probability model consists of an **experiment** which produces exactly one out of several mutually exclusive **outcomes**. The essential elements are:

1. The **sample space** Ω. This is simply the collection of all possible outcomes.
2. A **probability law**, which assigns a "likelihood" to different **events**. More on this later.

An **event** A is simply a collection of possible outcomes, i.e., A is a subset of Ω. We denote the probability that A occurs as

$$\mathrm{P}(A).$$

The probability law $\mathrm{P}(\cdot)$ must obey certain properties, which we will get to soon, but first let's look at two simple examples.

Example. Consider a fair six-died die. The experiment is rolling the die. The sample space (possible outcomes) is given by

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Note that this definition of Ω involves a certain amount of idealization. In particular, we omit possibilities such as "the die balances perfectly on a corner or edge" and "the die rolls off the table". The determination of the sample space often involves such judgement calls.

Under this definition of Ω, events include (but of course are not limited to)

- $\{1\}$, i.e., the result of the roll is a "1"
- $\{1, 3, 5\}$, i.e., the result is odd
- $\{2, 4, 6\}$, i.e., the result is even
- $\{1, 2, 3\}$, i.e., the result is ≤ 3
- etc.

In this case, there are $2^6 = 64$ different possible events under the assumption that we allow $A = \emptyset$ and $A = \Omega$, which can be interpreted as the events that "nothing happens" and that "something happens", to quality as events. This might seem a little strange, but both $\emptyset$ and Ω are subsets of Ω, and we will shortly see that there are good reasons for letting them count as events.

Since the die is "fair", a natural probability law is to assign each of the six possible outcomes the same value, i.e.,

$$\mathrm{P}(\{1\}) = \mathrm{P}(\{2\}) = \mathrm{P}(\{3\}) = \mathrm{P}(\{4\}) = \mathrm{P}(\{5\}) = \mathrm{P}(\{6\}) = \frac{1}{6}.$$

It is then straightforward to compute the corresponding probability of different events, e.g.,

- $\mathrm{P}(\{1, 3, 5\}) = \frac{1}{2}$
- $\mathrm{P}(\{1, 2, 3\}) = \frac{1}{2}$
- $\mathrm{P}(\{1, 2\}) = \frac{1}{3}$
- etc.

Example. Consider a fair coin that we will toss twice. When we have repeated actions like this, we will often considered a single experiment with sample space; here we have

$$\Omega = \{HH, HT, TH, TT\}.$$

Events include

- $\{HH, HT, TH\}$, i.e., there is at least one 'heads' (or at most one 'tails')
- $\{HT, TH\}$, i.e., there is exactly one 'heads' (or exactly one 'tails')
- etc.

Since the coin is fair, a natural probability law is to assign each of the four events a probability of 1/4, and so

$$\begin{aligned} \mathrm{P}\left(\text{at least one 'heads'}\right) &= \mathrm{P}\left(\{HH, HT, TH\}\right) = 3/4, \\ \mathrm{P}\left(\text{at most one 'heads'}\right) &= \mathrm{P}\left(\{HT, TH\}\right) = 1/2. \\ &\text{etc.} \end{aligned}$$

Kolmogorov's probability axioms

We will build up a theory of probability based axioms that all probability laws must obey in order to be consistent with common sense. This abstraction allows us to develop definitive mathematical rules that stand apart from the philosophical questions about what the probability really represents. Specifically, we will require a probability law to assign a number to every possible event A such that

1. **Nonnegativity:** $\mathrm{P}(A) \geq 0$ for every event A
2. **Additivity:** If A and B are **disjoint**, i.e., if $A \cap B = \emptyset$, then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.
3. **Normalization:** $\mathrm{P}(\Omega) = 1$, that is, the probability that "something happens" is 1.

There are many properties that can be immediately derived from these three axioms. For example, the normalization and additivity axioms tell us that

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega \cup \emptyset) = \mathrm{P}(\Omega) + P\emptyset = 1 + \mathrm{P}(\emptyset).$$

Combining this with the nonnegativity axiom we have that

$$\mathrm{P}(\emptyset) = 0,$$

i.e., the probability that "nothing happens" is 0. Also, for any event A,

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c),$$

and so

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A).$$

Another useful property that follows from additivity is that if A_1, $A_2, \ldots, A_n$ are n disjoint events, then

$$\begin{aligned}
\mathrm{P}\left(A_1 \cup A_2 \cup \cdots \cup A_n\right) &= \mathrm{P}\left(A_1\right) + \mathrm{P}\left(A_2 \cup A_3 \cup \cdots \cup A_n\right) \\
&= \mathrm{P}\left(A_1\right) + \mathrm{P}\left(A_2\right) + \mathrm{P}\left(A_3 \cup \cdots \cup A_n\right) \\
&\vdots \\
&= \mathrm{P}\left(A_1\right) + \mathrm{P}\left(A_2\right) + \cdots + \mathrm{P}\left(A_n\right).
\end{aligned}$$

Here are some additional properties that you should prove at home. Proving these will help provide a good review of basic set theory as well.

Let A, B, C be arbitrary events and let $\mathrm{P}(\cdot)$ be a probability law satisfying the Kolmogorov axioms. Then

1. If $A \subseteq B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$
2. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
3. $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$
4. $\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(A^c \cap B) + \mathrm{P}(A^c \cap B^c \cap C)$

Exercise: Out of the students in a class, 60% love soda, 70% love pizza, and 40% love both soda and pizza. What is the probability that a randomly selected student loves neither soda nor pizza?

Where do probability laws come from?

That is a good question, and this is where the "modeling" comes in. A probability law can potentially be based on factors such as:

- relative frequencies in past occurrences (i.e., "data driven")
- physical laws
- subjective belief based on experience
- a careful and thorough polling of the public
- etc.

Examples where these different approaches can be exploited include:

- "What is the probability that LeBron James makes his next free throw?"
- "What is the probability that more than 10^3 photons hit the detector in 1 μs?"
- "What is the probability that my wife will be mad at me when I get home?"
- "What is the probability that Hillary Clinton will be elected president in 2016?"

Discrete models vs. continuous models

When there are a finite number of possible outcomes in Ω, defining all of the possible events does not require too much imagination. If

$$|\Omega| = n,$$

where $|\Omega|$ denotes the size or number of elements in Ω, then there are 2^n different subsets.

In many situations Ω can be *huge* but still easy to describe:

- The number of 13-card bridge hands you could be dealt
- The number of possible license plates you could potentially receive
- The number of possible outcomes for all teams for the entirety of one Major League Baseball season

Here, the probability of an event is simply the sum of the probabilities of the outcomes that make up that event. Thus, if $A = \{s_1, s_2, \ldots, s_m\}$ then

$$\mathrm{P}(A) = \mathrm{P}(s_1) + \mathrm{P}(s_2) + \cdots + \mathrm{P}(s_m).$$

Moving from a finite number of discrete events to an **infinite number** of **discrete events** doesn't cause too many mathematical difficulties.

Example. You flip coin until you see "tails". The outcome of the experiment is how many times the coin gets flipped. This could be any natural number, i.e.,

$$\Omega = \{1, 2, 3, \ldots\} = \mathbb{N}.$$

If the coin is "fair", a natural probability law is[1]

$$\mathrm{P}\left(k\right) = \mathrm{P}\left(\{k \text{ flips until "tails"}\}\right) = \left(\frac{1}{2}\right)^k.$$

It is easy to check that

$$\sum_{k=1}^{\infty} P(k) = 1.$$

In contrast to the discrete case, when there is a **continuum** of possible outcomes ("uncountably infinite" in the language of set theory[2]), then there are some very technical considerations about what subsets of Ω can constitute an event.

[1]Note that we are slightly abusing notation here by letting k stand for both "the number of flips until 'tails"' as well as "the event that there are k flips until 'tails"'.

[2]See the wikipedia page on Cantor's "diagonal argument" for a more complete understanding of the difference between "countably infinite" and "uncountably infinite".

For example, suppose I choose a point at random from the interval $[0, 1]$. The natural probability law would define the **probability of any particular point** to be **zero**. After all, what are the chances that you would draw $\frac{\sqrt{2}}{2} = 0.70710678118\ldots$ or $\frac{1}{5} = 0.20000000000\ldots$ ***exactly***? However, if I define the event A to be picking a point between $\frac{1}{3}$ and $\frac{2}{3}$, i.e., $A = [\frac{1}{3}, \frac{2}{3}]$, then in this case

$$\mathrm{P}\left(A\right) = \mathrm{Length}(A) = \frac{1}{3},$$

and similarly for any other "typical" subset $A \subseteq \Omega$.

However, there are some subsets for which the "length" of the subset is not well-defined—these are called "non-measurable sets". I'd give you an example, but it's not really worth it—these sets are so unusual that they rarely (if ever) play a role in our understanding of probability.

This issue, although seemingly arcane, is important to resolve to put probability on a firm mathematical footing. Fortunately, this has been done in an area of mathematics called "measure theory". This is a topic for first-year graduate students—in this class it is enough to know that assigning probabilities to well-defined subsets is enough to avoid any major difficulties.

The discrete uniform law

The most basic probability law is simply that every outcome has the same probability. If Ω is finite with $|\Omega| = n$, this simply means that for any $A \subseteq \Omega$,

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

Example. A fair six-sided die is rolled; call the outcome D. What is $\mathrm{P}(D < 5)$?

In this case, $A = \{1, 2, 3, 4\}$, and so $\mathrm{P}(A) = \frac{4}{6} = \frac{2}{3}$.

Example. A fair coin is flipped three times. What is the probability that exactly two "heads" occur?

In this case we are actually dealing with a sequence of outcomes. We will talk more about ways to handle such problems later on, but in this case we can simply expand our notion of the sample space Ω to include all possible sequences of outcomes, i.e., we can consider

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}.$$

Since Ω contains eight possible outcomes and each has equal probability (assuming a fair coin), each of these outcomes has a probability of $\frac{1}{8}$. There are only three outcomes that have exactly two heads:

$$A = \{\text{HHT, HTH, THH}\},$$

and so

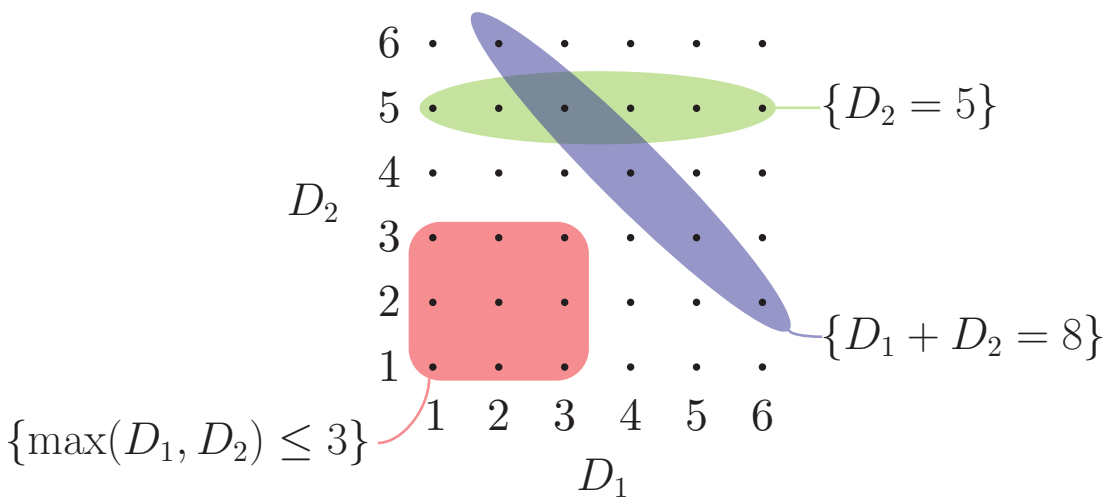
$$\mathrm{P}(A) = \frac{3}{8}.$$

Exercise: Calculate the probability of these events:

1. *at least* two heads
2. an odd number of heads
3. all tails

Exercise: We roll two fair six-sided dice; call the outcomes D_1 and D_2. There are now $6^2 = 36$ possibilities, each with equal probability.

Here is a graphical depiction of some events:



Calculate the probability that

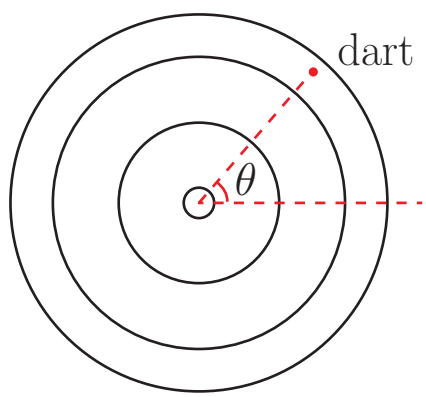
1. the first roll is larger than the second, i.e., $\mathrm{P}\left(D_1 > D_2\right)$
2. the first roll is equal to half of the second, i.e., $\mathrm{P}\left(D_1 = \frac{1}{2}D_2\right)$
3. at least one roll is a four, i.e., $\mathrm{P}\left(\{D_1 = 4\} \cup \{D_2 = 4\}\right)$

The continuous uniform law

When Ω is a continuum of events, the statement "every outcome is equally likely" becomes trickier, since the outcome of any particular event is zero.

In many cases, it will be natural to take Ω as an interval on the real line $\mathbb{R}$, or as a subset of the plane $\mathbb{R}^2$, or as a subspace of the space $\mathbb{R}^3$, etc.

For example, suppose I throw a dart at a dartboard and ask what angle (in radians) the result makes with respect to the x-axis.



In this case, we can take $\Omega = [0, 2\pi]$ (or $[-\pi, \pi]$).

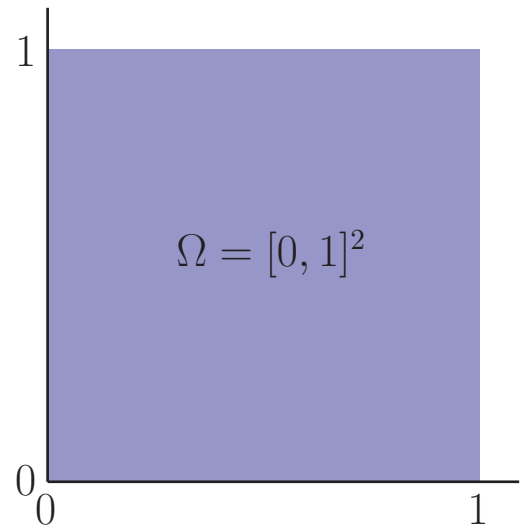
Then events A are subsets of Ω, and the uniform law assigns

$$\mathrm{P}(A) = \frac{\mathrm{Length}(A)}{\mathrm{Length}(\Omega)}.$$

In the dartboard example

$$\mathrm{P}\left(\frac{\pi}{4} \leq \theta \leq \frac{\pi}{2}\right) = \frac{\frac{\pi}{4}}{2\pi} = \frac{1}{8}.$$

Example. Suppose Ω is the **unit-square** $[0,1]^2 = [0,1] \times [0,1]$, i.e., $\Omega = \{(x,y) : 0 \le x \le 1, 0 \le y \le 1\}$.



Then events A are subsets of $[0,1]^2$ and

$$\mathrm{P}(A) = \frac{\mathrm{Area}(A)}{\mathrm{Area}(\Omega)} = \mathrm{Area}(A).$$

Exercise: With $\Omega = [0,1]^2$ and $A = \{\max(x,y) \le \frac{1}{3}\}$. What is $\mathrm{P}(A)$?

Exercise: Han and Chewbacca have arranged to meet at the cantina at noon. Unfortunately Han gets delayed by a bounty hunter and Chewbacca loses his watch, so they both are running late. Suppose that they both arrive with delays of anywhere from zero to two hours (with all possible delay combinations equally likely). Whoever gets there first will have a drink, wait for 20 minutes, and will leave if the other has not yet arrived. What is the probability that Han and Chewbacca meet? (Hint: start by sketching the event A.)

Background and Review: Basic set operations

As we have seen in this set of notes, it is very natural to talk about sample spaces, outcomes and events in terms of set operations. This section serves as a quick brush-up on the basics[3].

A **set** is just a collection of objects. For example

$$D = \{\text{Clinton, Biden, Warren}\},$$
$$R = \{\text{Rubio, Cruz, Christie}\},$$
$$\mathbb{Z}_8 = \{0, 1, 2, 3, 4, 5, 6, 7\},$$

are examples of *finite* sets, wheras the following

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, 3, \ldots\}$$
$$\mathbb{Q} = \left\{ \frac{a}{b} \text{ for all } a, b \in \mathbb{Z} \right\},$$

are examples of *countably infinite* sets. Finally, sets like

$$\mathbb{R} = \{\text{all real numbers}\}$$
$$U = \{x \in \mathbb{R} \ : \ 0 \leq x \leq 1\},$$

are examples of *uncountably infinite* sets.

The so-called **empty set** $\emptyset$ is the set which contains nothing,

$$\emptyset = \{\ \}.$$

[3] The Wikipedia page http://en.wikipedia.org/wiki/Set_operations_(Boolean) has a more extensive review. So does the required text.

A set B is a **subset** of another set A if everything in B is also in A:

$$B \subset A, \quad \text{if and only if for every } x \in B \text{ we also have } x \in A.$$

The empty set $\emptyset$ is a subset of every set.

For everything we do in this class, all sets of interest will be subsets of a **sample space** Ω — you can think of Ω as the "universe" associated with a particular experiment.

Example: Suppose that A is a finite set with n elements.

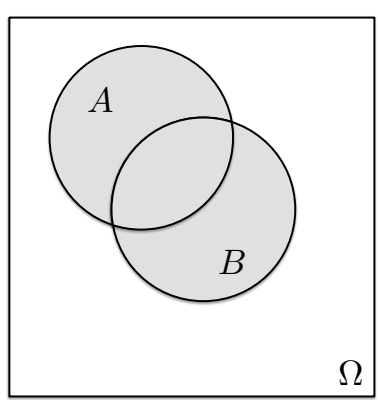
1. How many subsets of A have exactly one element?
2. How many subsets of A have exactly two elements?
3. How many subsets of A are there total?

Set operations

Union: Simply combine the elements of the two sets. Easy example:

$$\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}.$$

Here is a picture:

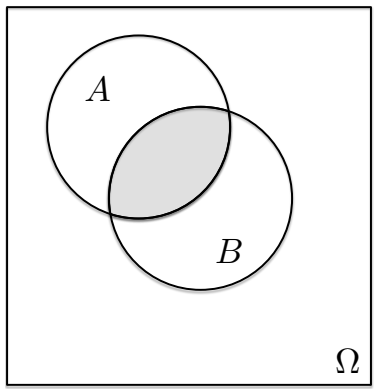


$A \cup B$ shaded

Intersection: Find the common elements between two sets. Easy example:

$$\{1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}.$$

Here is a picture:

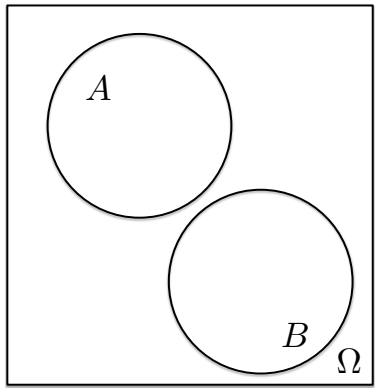


$A \cap B$ shaded

We say that A and B are **disjoint** or **mutitally exclusive** is they have no elements in common,

$$A \cap B = \emptyset.$$

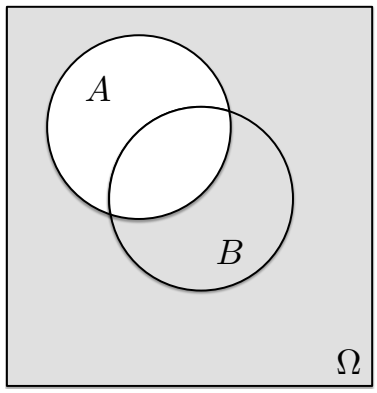
Here is a picutre:



The **complement** A^c of A is everything in Ω that is not in A. Easy example:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad A = \{1, 2\}, \quad A^c = \{3, 4, 5, 6\}.$$

Here is a picture:

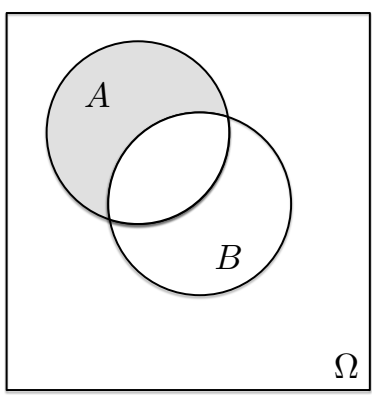


A^c shaded

The **difference** between A and B is everything in A which is not in B,

$$A \backslash B = A \cap B^c.$$

Here is a picture:



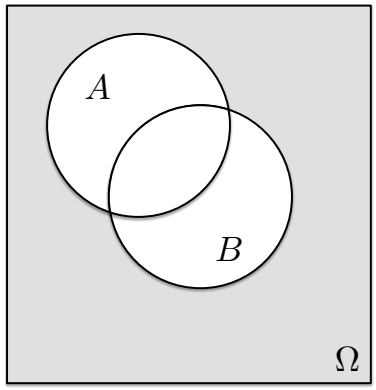
$A \backslash B$ shaded

Obviously, $A \backslash B \neq B \backslash A$ in general.

DeMorgan's Laws

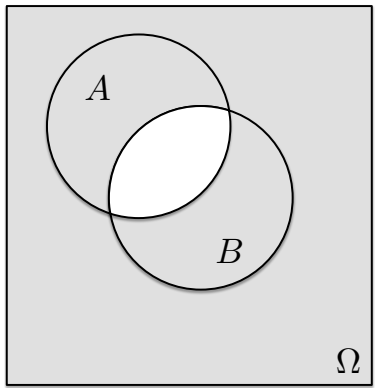
Two simple rules of set algebra come in handy from time to time.

1. $(A \cup B)^c = A^c \cap B^c$. Here is a picture:



$(A \cup B)^c$ shaded

2. $(A \cap B)^c = A^c \cup B^c$. Here is a picture:



$(A \cap B)^c$ shaded

Exercise: Suppose our sample space is $\Omega = \{2, 4, 6, 8, 10, 12\}$, and let

$$A = \{2, 4, 6\}, \quad B = \{8, 10, 12\}.$$

Find

1. $A \cup B$
2. $B \cup C$
3. $A \cap C$
4. $(B \cup C)^c$
5. $(A \cup (B^c \cup (B \cap C)^c))^c$