Sample space, events

The general setting is: We perform an experiment which can have a number of different outcomes. The *sample space* is the set of all possible outcomes of the experiment. We usually call it $\mathcal{S}$.

It is important to be able to list the outcomes clearly. For example, if I plant ten bean seeds and count the number that germinate, the sample space is

$$\mathcal{S} = \{0,1,2,3,4,5,6,7,8,9,10\}.$$

(This notation means the set whose members are $0,1,\ldots,9,10$. See the handout on mathematical notation.)

If I toss a coin three times and record the results of the three tosses, the sample space is

$$\mathcal{S} = \{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT\},$$

where (for example) HTH means 'heads on the first toss, then tails, then heads again'.

If I toss a fair coin three times and simply count the number of heads obtained, this is a different experiment, with a different sample space, namely $\mathcal{S} = \{0,1,2,3\}$.

Sometimes we can assume that *all the outcomes are equally likely*. Don't assume this unless either you are told to, or there is some physical reason for assuming it. In the beans example, it is most unlikely. In the coins example, the assumption will hold if the coin is 'fair': this means that there is no physical reason for it to favour one side over the other. Of course, in the last example, the four outcomes would definitely not be equally likely!

If all outcomes are equally likely, then each has probability $1/|\mathcal{S}|$. (Remember that $|\mathcal{S}|$ is the number of elements in the set $\mathcal{S}$.)

On this point, Albert Einstein wrote, in his 1905 paper *On a heuristic point of view concerning the production and transformation of light* (for which he was awarded the Nobel Prize),

In calculating entropy by molecular-theoretic methods, the word "probability" is often used in a sense differing from the way the word is defined in probability theory. In particular, "cases of equal probability" are often hypothetically stipulated when the theoretical methods employed are definite enough to permit a deduction rather than a stipulation.

In other words: Don't just assume that all outcomes are equally likely, *especially* when you are given enough information to calculate their probabilities!

An *event* is a subset of $\mathcal{S}$. We can specify an event by listing all the outcomes that make it up. In the above example, let A be the event 'more heads than tails' and B the event 'tails on last throw'. Then

$$
\begin{aligned}
A &= \{HHH, HHT, HTH, THH\}, \\
B &= \{HHT, HTT, THT, TTT\}.
\end{aligned}
$$

The probability of an event is calculated by adding up the probabilities of all the outcomes comprising that event. So, *if all outcomes are equally likely*, we have

$$
P(A) = \frac{|A|}{|\mathcal{S}|}.
$$

In our example, both A and B have probability $4/8 = 1/2$.

An event is *simple* if it consists of just a single outcome, and is *compound* otherwise. In the example, A and B are compound events, while the event 'heads on every throw' is simple (as a set, it is $\{HHH\}$). If $A = \{a\}$ is a simple event, then the probability of A is just the probability of the outcome a, and we usually write $P(a)$, which is simpler to write than $P(\{a\})$. (Note that a is an *outcome*, while $\{a\}$ is an *event*, indeed a simple event.)

We can build new events from old ones:

- $A \cup B$ (read 'A union B') consists of all the outcomes in A or in B (or both!);

- $A \cap B$ (read 'A intersection B') consists of all the outcomes in both A and B;

- $A \setminus B$ (read 'A minus B') consists of all the outcomes in A but not in B;

- A' (read 'complement of A') consists of all outcomes not in A (that is, $\mathcal{S} \setminus A$);

- $\emptyset$ (read 'empty set') for the event which doesn't contain any outcomes.

Note the backward-sloping slash; this is not the same as either a vertical slash $|$ or a forward slash $/$. We also write

- $A \subseteq B$ (read 'A is a subset of B' or 'A is contained in B') to show that every outcome in A is also in B.

In the example, A' is the event 'at least as many tails as heads', and $A \cap B$ is the event $\{HHT\}$. Note that $P(A \cap B) = 1/8$; this is not equal to $P(A) \cdot P(B)$, despite what you read in some books!

What is probability?

Nobody really knows the answer to this question.

Some people think of it as 'limiting frequency'. That is, to say that the probability of getting heads when a coin is tossed means that, if the coin is tossed many times, it is likely to come down heads about half the time. But if you toss a coin 1000 times, you are not likely to get exactly 500 heads. You wouldn't be surprised to get only 495. But what about 450, or 100?

Some people would say that you can work out probability by physical arguments, like the one we used for a fair coin. But this argument doesn't work in all cases (for example, if one side of the coin is made of lead and the other side of aluminium); and it doesn't explain what probability means.

Some people say it is subjective. You say that the probability of heads in a coin toss is $1/2$ because you have no reason for thinking either heads or tails more likely; you might change your view if you knew that the owner of the coin was a magician or a con man. But we can't build a theory on something subjective.

We regard probability as a way of modelling uncertainty. We start off with a mathematical construction satisfying some axioms (devised by the Russian mathematician A. N. Kolmogorov). We develop ways of doing calculations with probability, so that (for example) we can calculate how unlikely it is to get 480 or fewer heads in 1000 tosses of a fair coin. The answer agrees well with experiment.

Kolmogorov's Axioms

Remember that an event is a subset of the sample space S. Two events A and B are called *disjoint* if $A \cap B = \emptyset$, that is, no outcome is contained in both. A number of events, say $A_1, A_2, \ldots$, are called *mutually disjoint* or *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for any two of the events A_i and A_j; that is, no two of the events overlap.

According to Kolmogorov's axioms, each event A has a probability $P(A)$, which is a number. These numbers satisfy three axioms:

Axiom 1: For any event A, we have $P(A) \geq 0$.

Axiom 2: $P(S) = 1$.

Axiom 3: If the events $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

Note that in Axiom 3, we have the union of events and the sum of numbers. Don't mix these up; *never* write $P(A_1) \cup P(A_2)$, for example. Sometimes we separate Axiom 3 into two parts: Axiom 3a if there are only finitely many events $A_1, A_2, \ldots, A_n$, so that we have

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i),$$

and Axiom 3b for infinitely many. We will only use Axiom 3a, but 3b is important later on.

Notice that we write

$$\sum_{i=1}^{n} P(A_i)$$

for

$$P(A_1) + P(A_2) + \cdots + P(A_n).$$

Proving things from the axioms

You can prove simple properties of probability from the axioms. That means that every step must be justified by appealing to an axiom. These properties seem obvious, just as obvious as the axioms; but the point of this game is that we assume only the axioms, and build everything else from that.

Here are some examples of things proved from the axioms. There is really no difference between a theorem, a proposition, a lemma, and a corollary; they are all fancy words used by mathematicians for things that have to be proved. Usually, a theorem is a big, important statement; a proposition a rather smaller statement; a lemma is a stepping-stone on the way to a theorem; and a corollary is something that follows quite easily from a theorem or proposition that came before.

Proposition 1 $P(A') = 1 - P(A)$ for any event A.

Proof Let $A_1 = A$ and $A_2 = A'$ (the complement of A). Then $A_1 \cap A_2 = \emptyset$ (that is, the events A_1 and A_2 are disjoint), and $A_1 \cup A_2 = \mathcal{S}$. So

$$
\begin{aligned}
P(A_1) + P(A_2) &= P(A_1 \cup A_2) \quad \text{(Axiom 3)} \\
&= P(\mathcal{S}) \\
&= 1 \quad \quad \quad \quad \text{(Axiom 2).}
\end{aligned}
$$

So $P(A') = P(A_2) = 1 - P(A_1) = 1 - P(A)$. ∎

Once we have proved something, we can use it on the same basis as an axiom to prove further facts.

Corollary $P(A) \leq 1$ for any event A.

Proof We have $1 - P(A) = P(A')$ by Proposition 1, and $P(A') \geq 0$ by Axiom 1; so $1 - P(A) \geq 0$, from which we get $P(A) \leq 1$. ∎

Remember that if you ever calculate a probability to be less than 0 or more than 1, you have made a mistake!

Corollary $P(\emptyset) = 0$.

Proof For $\emptyset = \mathcal{S}'$, so $P(\emptyset) = 1 - P(\mathcal{S})$ by Proposition 1; and $P(\mathcal{S}) = 1$ by Axiom 2, so $P(\emptyset) = 0$. ∎

Here is another result.

Proposition 2 If $A \subseteq B$, then $P(A) \leq P(B)$.

Proof Take $A_1 = A$, $A_2 = B \setminus A$. Again we have $A_1 \cap A_2 = \emptyset$ (since the elements of $B \setminus A$ are, by definition, not in A), and $A_1 \cup A_2 = B$. So by Axiom 3,

$$P(A_1) + P(A_2) = P(A_1 \cup A_2) = P(B).$$

In other words, $P(A) + P(B \setminus A) = P(B)$. Now $P(B \setminus A) \geq 0$ by Axiom 1; so

$$P(A) \leq P(B),$$

as we had to show. ∎

Finite number of outcomes

Although the *outcome x* is not the same as the *event* $\{x\}$, we are often lazy and write $P(x)$ where we should write $P(\{x\})$.

Proposition 3 If the event B contains only a finite number of outcomes, say $B = \{b_1, b_2, \ldots, b_n\}$, then

$$P(B) = P(b_1) + P(b_2) + \cdots + P(b_n).$$

Proof To prove the proposition, we define a new event A_i containing only the outcome b_i, that is, $A_i = \{b_i\}$, for $i = 1, \ldots, n$. Then $A_1, \ldots, A_n$ are pairwise disjoint (each contains only one element, which is in none of the others), and $A_1 \cup A_2 \cup \cdots \cup A_n = B$; so by Axiom 3, we have

$$P(B) = P(A_1) + P(A_2) + \cdots + P(A_n) = P(b_1) + P(b_2) + \cdots + P(b_n). \quad \blacksquare$$

Corollary If the sample space $\mathcal{S}$ is finite, say $\mathcal{S} = \{s_1, \ldots, s_n\}$, then

$$P(s_1) + P(s_2) + \cdots + P(s_n) = 1.$$

Proof For $P(s_1) + P(s_2) + \cdots + P(s_n) = P(\mathcal{S})$ by Proposition 3, and $P(\mathcal{S}) = 1$ by Axiom 2. $\quad \blacksquare$

Now we see that, if all the n outcomes are equally likely, and their probabilities sum to 1, then each has probability $1/n$, that is, $1/|\mathcal{S}|$. Now going back to Proposition 3, we see that, *if all outcomes are equally likely*, then

$$P(A) = \frac{|A|}{|\mathcal{S}|}$$

for any event A, justifying the principle we used earlier.

Inclusion-Exclusion Principle
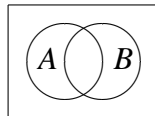
A Venn diagram for two sets A and B suggests that, to find the size of $A \cup B$, we add the size of A and the size of B, but then we have included the size of $A \cap B$ twice, so we have to take it off. In terms of probability:

Proposition 4 For any two events A and B:

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

Proof We now prove this from the axioms, using the Venn diagram as a guide. We see that $A \cup B$ is made up of three parts, namely

$$A_1 = A \cap B, \qquad A_2 = A \setminus B, \qquad A_3 = B \setminus A.$$

Indeed we do have $A \cup B = A_1 \cup A_2 \cup A_3$, since anything in $A \cup B$ is in both these sets or just the first or just the second. Similarly we have $A_1 \cup A_2 = A$ and $A_1 \cup A_3 = B$.

The sets A_1, A_2, A_3 are pairwise disjoint. (We have three pairs of sets to check. Now $A_1 \cap A_2 = \emptyset$, since all elements of A_1 belong to B but no elements of A_2 do. The arguments for the other two pairs are similar – you should do them yourself.)
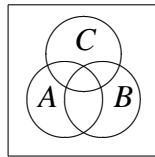
So, by Axiom 3, we have

$$\begin{aligned}
P(A) &= P(A_1) + P(A_2), \\
P(B) &= P(A_1) + P(A_3), \\
P(A \cup B) &= P(A_1) + P(A_2) + P(A_3).
\end{aligned}$$

Substituting from the first two equations into the third, we obtain

$$\begin{aligned}
P(A \cup B) &= P(A_1) + (P(A) - P(A_1)) + (P(B) - P(A_1)) \\
&= P(A) + P(B) - P(A_1) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}$$

as required. ∎

The Inclusion-Exclusion Principle extends to more than two events, but gets more complicated. Here it is for three events.

To calculate $P(A \cup B \cup C)$, we first add up $P(A)$, $P(B)$, and $P(C)$. The parts in common have been counted twice, so we subtract $P(A \cap B)$, $P(A \cap C)$ and $P(B \cap C)$. But then we find that the outcomes lying in all three sets have been taken off completely, so must be put back, that is, we add $P(A \cap B \cap C)$.

Proposition 5 For any three events A, B, C, we have

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Proof Put $D = A \cup B$. Then Proposition 4 shows that

$$\begin{aligned}
P(A \cup B \cup C) &= P(D \cup C) \\
&= P(D) + P(C) - P(D \cap C) \\
&= P(A \cup B) + P(C) - P(D \cap C) \\
&= P(A) + P(B) - P(A \cap B) + P(C) - P(D \cap C),
\end{aligned}$$

using Proposition 4 again.

The distributive law (see below, and also the Phrasebook) tells us that

$$D \cap C = (A \cup B) \cap C = (A \cap C) \cup (B \cap C),$$

so yet another use of Proposition 4 gives

$$
\begin{aligned}
P(D \cap C) &= P((A \cap C) \cup (B \cap C)) \\
&= P(A \cap C) + P(B \cap C) - P((A \cap B) \cap (B \cap C)) \\
&= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C).
\end{aligned}
$$

Substituting this into the previous expression for $P(A \cup B \cup C)$ gives the result. ∎

Can you extend this to any number of events?

Other results about sets

There are other standard results about sets which are often useful in probability theory. Here are some examples.

Proposition Let A, B, C be subsets of $\mathcal{S}$.

Distributive laws: $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ and
$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.

De Morgan's Laws: $(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$.

We will not give formal proofs of these. You should draw Venn diagrams and convince yourself that they work. You should also look in the Phrasebook.